Georgetown University
Occasional Papers on Machine Translation, No. 30
Leon E. Dostert, Director
R. R. Macdonald, Editor

General Report, 1952-1963

prepared by

R. R. Macdonald

Preface by

L. E. Dostert

Georgetown University Machine Translation Research Project
Washington, D. C.
June 1963

The research on which this report is based is the work of all members of the project, both past and present. The present members of the project (March 31, 1963) are:

Director:	Dostert, L. E.
Staff:	Boldyreff, Antonina
	Brown, A. F. R.
	Chaloupka, Bedrich
	Chennault, Anna
	Dekonsky, Katherina
	Kalikin, Eugen
	Macdonald, R. Ross
	Moyne, John A.
	Pacak, Milos
	Henisz, Bozena
	Woyna, Adam
	Zarechnak, Michael
Consultants:	Joos, Martin
	Trager, George L.

The past members of the project are named in connection with the various papers to which they contributed. These papers are listed at the end of the report.

Please address all communications relative to this report to Leon E. Dostert, Director, 1330 New Hampshire Avenue, N.W., Washington, D.C.

The research described in this Report has been made possible by grants of funds from the Central Intelligence Agency and from the National Science Foundation. These grants are listed here; in the case of the first four items, the National Science Foundation supplied one part of the funds, and the Central Intelligence Agency the other part, as indicated, but the combination was listed as a National Science Foundation Grant.

1.	National Science Foundation Central Intelligence Agency	G2723	$ 35,000 65,000	9/27/56
2.	National Science Foundation Central Intelligence Agency	G3867	35,000 90, 000	6/17/57
3.	National Science Foundation Central Intelligence Agency	G5513	36,000 150,000	6/ 6/58
4.	National Science Foundation Central Intelligence Agency	Supplement	 9,890	
5.	Central Intelligence Agency	XG2230	24,979	7/ 1/59
6.	Central Intelligence Agency	XG 2239	153,000	9/16/59
7.	Central Intelligence Agency	XG 2312	439,000	7/ 1/60
8.	Central Intelligence Agency	XG 2427	438,000	9/ 1/61
9.	Central Intelligence Agency	Supplement	250,000	to 3/31/63
	Total		$1,728,239	

Free machine time for experiments was supplied on occasion by a number of agencies, of which the Department of Defence and the Atomic Energy Commission in the United States, and Euratom in the European Community were the largest contributors.

PREFACE

In publishing this General Report, the result of the collective effort of come forty persons and covering ten years of continuing and intensive work in linguistic research aimed at machine translation, the Georgetown University Machine Translation Research Project is presenting more than a series of ephemeral technical papers. This Monograph does not follow in form or content the usual periodic progress report to a sponsor, although it constitutes a general supplement to the several topical progress reports submitted to the project's former sponsor, the Central Intelligence Agency.(1)

No transcendant hypothetical concept is advanced here for bilingual or multilingual automatic transfer procedures, or for generalized monolingual automatic structural analysis. Concern for conciseness and relevancy precluded the detailed listing of the large corpus of Russian or Russian-English scientific articles from which the dictionary was primarily built and on which the general research was mostly focused. The report does not include a voluminous reproduction of analytical printouts or experimental translation outputs. It does not contain any symbolic general representation of one or another set of monolingual structural operations.

The analytical listings with coded data as well as the concordance or contextual output do not involve procedures of the type advanced for 'generative grammar'. Essentially, the analysis procedures do not aim at prescriptive 'rules' but rather at data-based or data-verified descriptive formulas.

When illustrative texts are given; they are not sentences or passages contrived to prove or invalidate any aprioristic tenets. This report makes no argument as to the feasibility or hopelessness of MT, or as to its economic practicality and scientific actual or potential usefulness. This belated concern among some has been outdistanced by the attainments of the recent past.

The report strives to be an orderly, lucid, accessible objective description of a decade of result-oriented experimental research, based on generally accepted theories and procedures of linguistic science, and focused on the gradual attainment of useful bilingual translation (primarily of Russian-English), in certain areas of science and technology.

---

(1) From the initial phase of the project, the Central Intelligence Agency, and, at the beginning only, the National Science Foundation, contributed a total of $1, 728, 239. 00, of which grateful acknowledgement is made. On March 31, 1963, this subsidy ceased, the position now being that the government-supported program for MT is to be co-ordinated among several government agencies through the National Science Foundation, which by statutory requirements supports only basic research.

Over ten years ago, first contacts with early workers in the field made clear the vastness and complexity of the problems involved in the automatic translation of natural languages, even if the transfer were to be one-directional and limited in scope to certain fields of discourse. At the beginning, too, the divergent assumptions, the variety of contemplated procedures (pre-editing, post-editing, simplified English, 'metalanguage', etc.) and the sometimes hazy and limitless hypotheses of some of the workers during the initial phase, led to the adoption of the experimental cumulative procedure at Georgetown. For that reason also, the project was oriented toward the early attainment of a partial, modest, but in our view, significant demonstration of feasibility. This first 'prise de position' was to lead to the publicly announced, and therefore somewhat distorted, Georgetown-IBM experiment in January 1954. (2)

This first demonstration of feasibility was to be subsequently assessed as having not much more than 'historical significance'.(3) Whatever the significance of the Georgetown-IBM experiment of 1954, it seems definitely established that research in the field has broadly expanded since that date.

Through the ten years covered, the Report shows that the experimental method, focused on actual texts and on the production of results, has been followed with reasonably gratifying results. (4)

Moreover, frequent computer tests and runs have been an integral part of the Georgetown procedure. We all know that many a promising paper program does not always produce the anticipated computer results.

No system is known to have produced continuous output texts that did not essentially follow the same basic procedure. The early hypothesis of the so-called word-for-word dictionary lookup procedure, producing non-discrete multilinear lexical outputs, and barren of structural procedures, has long since been abandoned. Likewise, the maximization of 'cluster' storage in vast memory

---

(2) See my report in Locke and Booth, Machine Translation of Languages. Wiley and Sons, New York, 1955.

(3) See E. Delavenay's popular treatment of the subject of MT in the series Que Sais-je, Paris, 1957. (It would not be without interest for a bibliographer to make a comparative study of publications on MT before and after the Georgetown-IBM experiment of 1954. See, for instance, E. and K. Delavenay, Bibliography of Mechanical Translation, Mouton & Co. , The Hague, 1960, and the much more complete recent listing by Josephine Walkowicz, A Bibliography of Foreign Developments in Machine Translation and Information Processing, National Bureau of Standards, Technical Note 193, U.S. Printing Office, Washington, D. C. , 1963.

(4) A forthcoming paper by Paul Garvin will deal with the concepts and methods underlying the 1954 experiment and their relevancy to later machine translation research and development.

computer prototypes does not seem, so far, to have yielded the expected results. Two other projects aiming at sample output were brought to follow procedures essentially similar to those of the Georgetown Project, the data, resources, and objectives being essentially the same.

It is to be regretted that an artificial dichotomy has gained currency between those whose efforts were described as 'theoretical' or 'basic' research, and others who sought to focus their research on current bilingual texts and attempted by cyclical procedures to produce increasingly acceptable outputs. The confusion thus created is evident in the record and the report of the Congressional Hearings conducted on the subject in 1960. (5)

In the case of the Georgetown Project, the principal sponsor's interest was less in computational language data processing than in the progressive improvement of a one-directional transfer program in two languages (Russian-to-English) in certain fields of science and technology.

It should be obvious that both the theoretical and experimental approaches are valid, and indeed ultimately convergent. If the research is in fact oriented toward machine translation (6), it matters little whether provisional and progressive results are sought experimentally on the computer in the course of the investigation, or whether paper research is pursued until presumably broader results can be attained. The notion that the 'empirical' approach is bound to end in a plateau of non-improvable output is highly conjectural. By the summer of 1964 an evaluation of that hypothesis will probably be feasible on the basis of experimental work and objective data and criteria.

---

(5) See the Report of the Committee on Science and Astronautics, U.S. House of Representatives, Eighty- sixth Congress, Second Session, Serial d. , June 28, 1960, Government Printing Office, Union Calendar No. 895, Home Report No. 2021.

(6) This is becoming an increasingly open question, as is attested by the proposal to change the name of the recently founded "Association for Machine Translation and Computational Linguistics" by eliminating 'Machine Translation' from the name altogether. In this connection it is pertinent to recall the first conclusions of the Congressional Hearings in May 1960, which seem to have had little impact on actual developments:

"The hearings on mechanical translation vividly pointed out the importance of a mechanical translation system to the overall intelligence and scientific effort of our Nation. With the advent of such a capability, a new approach will be taken by all segments of our culture to the reading of foreign documents. Truly, a capability of translation in reverse - that is, English into foreign languages - will open up new vistas and avenues for the exchange of cultural, economic, agricultural, technical and scientific documents that will present the American way of life to people throughout the world.

"The pursuit of this research and development program on mechanical translation is a must and should be vigorously continued to insure an early capability on a national effort." Cf. footnote No. 5.

Machine Translation Research at Georgetown has rested on and has been oriented by certain basic assumptions, procedures and objectives. It will be helpful to review them briefly.

The basic assumptions include the following:

1. The automatic translation of languages is essentially a problem of sign-substitutions, that is, formulation and programming of a substitution procedure permitting the signs of the target language to be so selected and arranged so as to convey the information contained in the signs of the source language.

2. The fundamental problem of machine translation is a linguistic one. It should therefore be primarily guided by the procedures accepted and current in linguistic investigations.

3. Natural language is not a mathematical or a logical structure and therefore it is at least open to question whether the term "mathematical linguistics" has actual validity. The fact that certain partial and stable operations in a language structure can be the subject of algorithmic formulations and symbolic representation should not create the illusion that we are dealing in fact with a procedure involving the rigorousness of mathematics.

4. The development of structuralism in contemporary linguistics is at the basis of the concept of machine translation, since, without structuration procedures, the idea of sign-substitutions or automatic transfer of linguistic data would hardly be conceivable.

5. Machine translation should accept for the foreseeable future a limited field of investigation. It should be obvious that the investigation of the totality of any given structure *is* a matter that will involve years of research, and is in fact never-ending Accepting more limited aims, a delineation of a more limited scope of investigation, and more precise objectives can permit advances toward increasing achievements.

6. Machine translation should focus on what is actual in the areas of the language with which it concerns itself. It should not be unduly concerned with what is 'possible' in language. We should be satisfied with the acceptance of the fact that an inherently ambiguous sentence in a given language cannot be accurately transferred by man or machine. One has in mind such contrived puzzlers as 'These men are revolting', or 'She made him a good husband because she made him a good wife'. Basically, the formulation of the transfer or substitution program must be based on formalism, in the sense that the categories, functions and relationships of the items of a language must be represented in a formalistic manner for MT transfer.

The working procedures of the project include the following basic operations:

1. The research is text-focused in the sense that the lexical buildup and the structural inventory is essentially, but not exclusively, text-derived.

2. The analysis of the structures involved is being carried out on the basis of actually encountered data and follows current linguistic procedures.

3. Word classes and grammar relationships, as traditionally established and modified by more recent structural procedures, have been retained.

4. The procedure defines the levels of investigation as the morphological (word level); the syntagmatic (word-group or phrase level) and the syntactic (sentence level).

5. The area of investigation which is not approached on the basis of the data established and available at the levels indicated under 4, has been called 'semantic categorization' in this project. This involves an attempt to discern formal cues for the resolution of lexical choice (polysemia) and other semantic problems.

As for the objectives sought by the project, they are essentially the following:

1. The progressive improvement of experimental runs by means of a feedback procedure on the basis of the discernment of lacunae and inadequacies.

2. The recognition of the fact that the language of science and technology is more accessible to programming for automatic transfer than other areas of language expression, and should therefore be undertaken first.

3. One-directional transfer should be sought in the initial phases of MT research.

4. Transfer from Russian to English is the primary goal.

As mentioned before, this General Report is the result of the individual and group efforts of members of the research staff. The work in Russian analysis was initially developed by Paul Garvin and later by Michael Zarechnak, and the fundamentals of the existing formulations in that field are essentially the results of Zarechnak's work. Their many co-workers included Milos Pacak, Bozena Henisz, Bedrich Chaloupka, Eugene Kalikin, Antonina Boldyreff, Philip Smith, to mention only a few whose work has been of significant value.

The research in English structure indispensable to a proper 'synthesis' Program had not received the necessary attention until primary responsibility for that task was assumed two years ago by R. Ross Macdonald, to whom we are grateful for the preparation of the present report. The consultant ship guidance on the part of William Austin. Martin Joos, George Trager, John DeFrancis, Lawrence Summers, Richard Harrell and Boris Unbegaum was most helpful and gladly acknowledged.

In programming, the early work of Peter Toma which led to the first significant continuous outputs for Russian to English, and that of his many assistants, is hereby recognized. The outstanding contribution of A. F. R. Brown in the development of programming procedures as well as the excellent work of John Moyne on the General Analysis Technique or Direct-Conversion system are gratefully acknowledged.

In closing a decade of work the Project now will proceed to continuing research under the sponsorship of the Atomic Energy Commission, the EURATOM Research Directorate and, very probably, the National Research Council of Canada.

To past and present sponsors grateful thanks are extended, since without their interest and support the results described in this General Report could not have been attained. To the other members of the staff who have contributed in various ways to this work, and who are not mentioned by name, I extend my personal thanks.

In closing, the careful, painstaking and orderly efforts of Dr. Macdonald in preparing, from a vast amount of separate papers, this General Report is acknowledged and sincerely appreciated.

Georgetown University L, E. Dostert
June 1963

TABLE OF CONTENTS

INTRODUCTION

This General Report covers the achievements of the Georgetown Machine Translation Project during the period from June 1952 until March 31, 1963.

The material comes from three sources.

1. In part, the report is a reworking of the Occasional Papers that were published during the period covered. There has been an attempt to render them more consistent in style and terminology, and more readable. Some of the more detailed sections have been summarized; some of the more elliptical sections have been expanded with those explanations which were clearly in the mind of the writer but which failed to flow from his pen; but no new material has been added without an indication of this fact in the text.

It must be pointed out that the papers on the work in Comparative Slavic are not described individually. First, only some of these papers have been published, and second, the scope of the research is such that it requires a special report of its own. Enough material is given, it is hoped, to allow the reader to judge of his interest in the forthcoming report on this phase of the research.

Similarly, the two systems of programming, the Direct Conversion and the SLC, are not described in detail. Each of these has been the subject of a comprehensive and fairly recent report, copies of either of which are still available. Brief descriptions of the systems are given in the last papers of this report, however, and sufficient other indications may be gleaned from passages throughout the text to allow the reader to judge of his interest.

A list of the Occasional Papers is given at the end of this General Report.

2. In part also, this report is a description of the present research at Georgetown; the results of this research will be published subsequently, but the trends are discussed in supplements to the appropriate papers.

3. In part again, this General Report embodies a number of terminal reports which were submitted to the Central Intelligence Agency on March 31, 1963, at the termination of their grants of support. Some of these terminal reports appear in this report as separate papers, since they cover material which had not been published at that time. Others of the terminal reports are incorporated with the Occasional Paper on the same topic. Thus, the section of this general report which deals with the dictionary begins with a description of the compilation of the dictionary (taken from the terminal report on the dictionary), continues with a description of the makeup of the dictionary (based on Occasional Paper No. 3), and ends with some statistics on the dictionaries (again from the terminal report) .

A list of the terminal reports if given at the end of the present publication

In its original form, this General Report contained a description of the manner in which translation is effected by machine so that those who are not: acquainted with the process might come to understand it. But the impossibility of being more than superficial in such a short space has made it seem advisable to leave the presentation of this information to other books which have such expectations as their chief purpose. One must hope that the description is clear enough that the reader will not experience major difficulties if he is unacquainted with the characteristics of a computer.

In the same way, a discussion of the linguistic attitudes which are evoked by the need of working with a machine has been put aside for want of adequate space. However, a brief description of the form of presentation of some of I linguistic statements seems advisable both because of the possibility that certain passages of the text might otherwise lack clarity, and because it will give at least some insight into the problems which are involved. This description follows the papers on Morphology, since it is at this point that it first becomes necessary.

In the reporting in general, greater emphasis has been placed on the principles involved, and less emphasis on the specific details of coding or classification. The reader who is interested in specific detail is referred to the original papers (which, though many are out of print, are available in various libraries) or is cordially invited to direct inquiries to the Director of the Georgetown Machine Translation Research Project.

THE HISTORY OF THE PROJECT

Machine translation has long been talked of in visionary terms. It was only during the Second World War, however, that it began to seem technologically possible. The first practical discussions of these possibilities began in 1945; they are described in the book: Machine Translation of Languages, edited by Locke and Booth.

The first interest in machine translation at Georgetown University was aroused in 1952.

In that year, in June, a formal meeting on machine translation was held at the Massachusetts Institute of Technology. Professor L. E. Dostert, Director of the Institute of Languages and Linguistics at Georgetown University, was invited to participate. Professor Dostert had worked in French-American liaison on the staff of General Eisenhower during the Second World War, and had been instrumental in devising and in implementing the system of simultaneous translation which was used at Nuremberg, and which is still used so largely in the United Nations and at many international conferences. He was very active in the instruction in simultaneous translation given at the Institute of Languages and Linguistics of Georgetown University, It was as a specialist in translation that he was asked to participate in the meeting at MIT. His paper was entitled: Ordinary Translation and Machine Translation.

Professor Dostert returned from the MIT meeting enthusiastic about the possibilities of machine translation. He began to speculate about it and to interest others in it. He consulted both other linguists and engineers and they gave the opinion that machine translation was feasible; a summation of these opinions is Dr. Paul Garvin's "Statement of Opinion Concerning Machine Translation" of April, 1953. The result of Professor Dostert's enthusiasm was an experiment in which Georgetown University and the International Business Machine Corporation collaborated.

The goal of this experiment was to produce an actual machine translation and to gain practical experience as to the problems involved in such an undertaking.

A number of sentences in the field of chemistry were selected. Linguistic analysis indicated that a translation could be effected with a dictionary of 250 words and a grammar of six syntactic operations. A transliteration of the Russian into the Latin alphabet was used in place of the Cyrillic alphabet in which Russian is conventionally written.

The words which were entered in the dictionary were divided into two classes; there were those which could usefully be analyzed into smaller components, and those which could not. Those which were analyzed were divided into a base and ending, and the bases and endings were entered separately in the dictionary. Each Russian word had one or two English equivalents in

the dictionary. Each Russian word had also certain diacritic signs or codes; there were three types of these.

One type of code indicated which of the six syntactic operations was to be used. Codes of this type were called program-initiating diacritics.

A second type of code indicated which of the English translations was to be chosen and also indicated what searches were to be made in deciding that choice. Codes of this type were called choice-determining diacritics.

The third type of code indicated the locations in the computer at which the information was stored which was to be used with the first two types of codes.

The six syntactic operations referred to by the program-initiating diacritics were numbered from 0 through 5 and were as follows:

0. The order of the original text is to be followed.

1. There is to be a difference of order in the translation from the order in the original, and an inversion is necessary.

2. There is a problem of choice; the choice depends on an indication which follows the word under consideration.

3. There is a problem of choice; the choice depends on an indication which precedes the word under consideration.

4. A word appearing in the original text is to be dropped, and no equivalents will appear in the translation.

5. At a point where there is no equivalent word in the original text, a word to be introduced into the translation.

After the dictionary and the program of operations had been prepared, a preliminary experiment was made to determine how they would function.

A group of people was asked to translate the sentences. They worked with the materials provided to the machine and as nearly as possible as the machine would work. The result was an acceptable English translation. Although the human beings were very accurate, they were rather slow, since they required an average of one minute for every word translated.

The dictionary, its codes, and the syntactic rules were keypunched and introduced into the machine's memory; (there were two types of memory, electrostatic core and magnetic drum).

Everything was ready at last for the final step.

Russian sentences were put into the machine.

The machine translated them into English.

On January 7, 1954, the results of the Georgetown-IBM experiment were announced in New York. The announcement was given wide publicity in the press and it elicited a wealth of comments from all quarters. Some people were astonished that machine translation was possible in any degree. Some people were disappointed because the translation was not as polished as English prose can be. But those who understood the difficulties of machine translation realized that a definite and decisive first step had been taken along a road that still wound far ahead before the final goal of excellent machine translation could be reached.

In a summation of the results of this experiment, Professor Dostert made six points.

1. The methods gives practical results: It produces an authentic machine translation from Russian to English.

2. The results do not indicate a need for either pre-editing of the input text nor for post-editing of the output text.

3. The problem of machine translation is primarily one of linguistic analysis and of the comparison of linguistic structures.

4. The ultimate basis for a systematic and widely useful coding of dictionary entries is the delimitation of features of meaning as well as the delimitation of features of structure.

5. The building up of technical dictionaries for various fields of discourse will undoubtedly be an effective procedure for solving certain problems of meaning.

6. There is the possibility of developing an artificial intermediate language system (a core language) by means of which translation between many pairs of languages will be facilitated.

Despite the public interest awakened by the Georgetown-IBM experiment, little official interest was aroused, and there was no official support for further research. However, the Georgetown team continued its work on a limited scale during the period from January 1954 until early 1956.

Early in 1956 the Institute of Precision Mechanics and Computer Technology of the U. S. S. R. Academy of Sciences announced a successful translation of English into Russian on their BESM Computer; they acknowledged the relationship between their undertaking and the Georgetown-IBM experiment, which they had followed with interest.

In June 1956, Georgetown University received a substantial grant from the National Science Foundation to undertake intensive research for the

translation of Russian scientific materials into English. And so, in the fall of 1956, a full-scale project with more than twenty research workers was organized, with L. E. Dostert as the director. The work was focused on the translation of Russian texts in the field of organic chemistry; this followed a precedent which has been set in the Georgetown-IBM experiment.

It was immediately realized that considerable training and orientation would be required to carry out the specific type of research assigned to each group and to assure proper co-ordination. The first three months were considered simply as investment in training; nonetheless, some useful results were obtained during this period.

From the outset, a weekly two-hour seminar was conducted. Papers were delivered and a system of seminar working papers was evolved. These papers were mimeographed and mailed out to those people who might be interested in them.

In order to keep the research workers fully aware of the situations which their linguistic formulations would encounter when subjected to programming processes, a programmer was assigned to the regular staff.

Two groups were organized, one for translation analysis, and one for linguistic analysis.

The translation analysis group concentrated on the preparation of a consistent translation.

The first requisite in the preparation of a dictionary was a man-made translation of the text. The commercial translation published by the Soviet Union was not considered to be sufficiently standardized for the purpose since it exhibited a great deal of elegant variation, that is, of the use of several terms in English to avoid the repetition of a term which was consistently the same in Russian. The translation analysis group prepared and standardized a new English translation so that each word in Russian had, where reasonably possible, one translation in English. However, the original Russian was modified neither in substance nor in structure. Only the English was affected by the selection of the guide translation and this selection was intended only to minimize the inconsistencies of human translation and to provide a consistent basis for the coding of the dictionary entries.

The linguistic analysis group concentrated on the analysis of the Russian language for machine translation; it subsequently came to be named the Experimental Group.

The work of the Experimental Group began with a detailed study of the Georgetown-IBM experiment of January 1954.

Certain decisions were taken quickly.

1 The system of transliteration was to be replaced by the Cyrillic alphabet.

2. No dictionary entries were to be split into base and ending.

3. The method of coding dictionary entries according to program-initiating diacritics and choice-determining diacritics was to be reworked and expanded.

4. The six original programs were to be recast and subdivided.

5. A corpus of eighty sentences was selected for study. Some of the problems found in this text were to be solved ad hoc; other existing problems were to be ignored. The purpose of the study was twofold: the researchers hoped to discover and to solve the problems of vocabulary selection and sentence arrangement; they also hoped to refine and to generalize as far as possible the translation operations employed in the original procedure. The first policy decision was that the various problems should be solved one at a time and solved completely. But very little work proved that the problems were so interrelated that any such piecemeal approach was impractical. It became necessary to devise an overall approach and a plan for the constant modification of what had been done to accommodate what had to be done. At this point in the discussions, unanimity gave way to a flood of opinions.

> In October 1956, another meeting on mechanical translation was held at the Massachusetts Institute of Technology. Among the participants were the University of Washington, the University of California in Los Angeles, the International Telemeter Corporation, Harvard University, Georgetown University, the Massachusetts Institute of Technology, the Cambridge Language Research Unit of England. The USSR was invited to send representatives but reported unable to do so.

When the Experimental Group had been working only a few months, it became evident that there was a considerable divergence of point of view. This divergence was fostered by Professor Dostert because he felt that the policy of giving each method a chance to show its mettle in free competition with the others and of then selecting whichever method was best adapted to machine translation would be more productive than a policy of predetermining a method and of forcing machine translation to conform to it. And so, by January of 1957, there were a number of groups where there had been only one in September, 1956. These groups acquired names derived from the names of the methods they advocated. The methods were known as code-matching, syntactic analysis, general analysis, and the sentence-by-sentence method. In brief, the methods have the following characteristics.

Code-matching begins with the coding of each Russian dictionary entry of each English meaning. The codes represent the various grammatical and associative functions which each word can fulfill. After the words of a Russian

text have been looked up in the dictionary and the codes copied, the analysis proceeds word-by-word through the text - always from left to right except for occasional regressions of one word to the left. The one function of the form which is applicable in each environment is arrived at by a comparison of the codes of contiguous words, those which are the same are selected and grouped the others are disregarded in that context. Certain modifications are necessary of course, when none of the codes of contiguous words match. The translation is effected by applying various mathematical processes to the strings of codes selected and to the codes of the English meanings. The proponent of this method was Miss Ariadne Lukjanow.

Syntactic analysis proceeds one sentence at a time. The machine analyses each sentence in terms of its immediate constituents, and then in terms of the immediate constituents of those immediate constituents, and so on until the analysis is complete. At each point the analysis concentrates on the item why conveys the largest amount of grammatical information. This item serves as a fulcrum in the process of prying out further information from the remaining item. For this reason, syntactic analysis is sometimes referred to as the fulcrum method. The proponent of this method was Dr. Paul Garvin.

General analysis also works with a sentence at a time. Each sentence is analysed into translation units whose presence, absence and positional relationship to each other are all important. The analysis is carried out at every possible level that will elicit useful information. Word-formation (morphology) is the first level. This also includes word-collocation (idiom). Word-grouping (syntagmatic processes) is the second level. This includes the agreement of adjectives with nouns, the government of nouns by verbs, or other form classes, and the modification of adjectives, verbs and other advert by adverbs. The organization of word-groups into sentences (syntax) is the third level. This is specifically the relationship of subject to predicate. The possibility that there are other levels is not precluded. The proponent of this method was Michael Zarechnak.

The sentence-by-sentence method is undoubtedly the most novel approach to machine translation attempted at Georgetown. Its proponent, Dr. A. F. Brown, proposed to develop a method of machine translation which the linguist could control by himself, and in which all of the information and processes , were readily available to the linguist whenever it was necessary to add, subtract or alter. Dr. Brown proposed to begin by translating one sentence of French into English, and then listing and filing all of the information and procedures needed to do this. Then he proposed to go on to a second sentence, and to revise and add to his file so that the adjusted system would translate both sentences, then to go on to a third, adding and revising, and to continue in this way until the system no longer required modifications or additions to handle new sentences.

The various groups differed not only in theory, but also in many details of practice. The matter of dictionary entries is an example. Split entries (in which each base is listed once and each ending which can be used with a

8

large number of different bases is also listed once) have the advantage of requiring less space in the memory, but the disadvantage of requiring more time, since it is necessary to look up first the base and then the ending and to collate them. Unsplit entries (in which each base is listed separately as many times as it has differing endings) have the advantage of saving time in the lookup procedure but the disadvantage of requiring a multitude of entries in the dictionary. In the Georgetown-IBM experiment, the entries were split when possible. In the work of the Experimental Group the entries were unsplit at first and then split. In the work of the later groups, various paths were followed. The syntactic analysis and the general analysis groups used split entries. The code-matching group used unsplit entries but stated that the method did not preclude the use of split entries. The sentence-by-sentence method resolved the dilemma by invoking a third factor, that of programming, and kept down the increase in both the size of the dictionary and the length of time required to look up a word by means of a different pattern of storage in the memory.

By March of 1957, it had become a deliberate policy to allow the broadest latitude to diversity of approach and of methods with the understanding that, when the various methods were tested, the one which responded best in a practical situation would be favored over the others.

At the beginning of April 1957, it was decided what the test should be. A selection of texts in the field of organic chemistry was chosen from the Soviet publication: The Journal of General Chemistry. The material might be studied beforehand to any degree desired, and then this 'prepared text' was to be translated into English. At the same time another passage, not previously selected or studied, was to be translated also. The translation of a prepared text had already been effected on other occasions, and so was not a novelty. But the translation of a 'random text' was felt to be the only really telling method of proving a translation system, and this had not yet been attempted.

The first step was the preparation of the dictionary. Some 24, 000 words were drawn from all issues of the publication beginning with the first issue of the year 1952. Because of repetitions in different contexts, only one quarter of these words were entered as individual lexical items. These six thousand individual lexical items were to be coded and analyzed by September 1957. This was felt to be an adequate dictionary for a first test run. It was established that the proportion of general vocabulary to chemical vocabulary in the text was as three to two. This fact emphasized the need for up-to-date text-based dictionaries in each technical field, and long-range plans were discussed for the collation of a number of dictionaries in various fields.

While the dictionary material was being selected, the various research groups continued to develop their various approaches.

On April 12 and 13, 1957, the Round Table Meeting on Machine Translation was held at Georgetown University. The detailed report is

to be found in: Report on the Eighth Annual Round Table Meeting on 1 Linguistics and Language Study: Research in Machine Translation; edited by Leon Dostert, Monograph No. 10, 1957.

During the summer of 1957, it was decided that a seminar on machine translation would be held as part of the Linguistic Institute **at** University of Michigan. This decision underlined the linguists' growing awareness of machine translation and of its problems.

In August 1957, the Eighth International Congress of Linguists met in Oslo. Dr. Paul Garvin represented Georgetown at this Congress and two papers on machine translation as developed at Georgetown wen presented. A report is to be found in the Proceedings of the VIII International Congress of Linguists; Oslo University Press, Oslo, Norway; 1958. Dr. Garvin also visited the Cambridge Language Research Unit in the United Kingdom and consulted with them for several days.

In September 1957, a number of decisions were taken by the Georgetown Machine Translation Research Project.

It was settled that the machine test would take place early in 1958. Each of the various methods that had been suggested for mechanical translation was to be given a fair trial and the best method selected for further development.

It was also decided that research on additional languages would be carried out. This research would not necessarily be aimed directly at machine translation from those languages, but at determining whether problems which arise in other languages might shed more light on the main task of machine translation from Russian into English. The languages selected were German, French, Arabic, and Chinese.

It was also foreseen that, by the end of 1957, a decision would be required as to the investigation of additional fields of scientific literature for machine translation, since the exclusive preoccupation with organic chemistry was providing a rather limited experience, especially in dictionary work.

It was decided that all four groups would continue their research along the current lines.

Because each group required more time to perfect its method, it was not possible to hold a full-scale test in early 1958. Only the general analysis group (Michael Zarechnak, Jane Pyne and others) was ready for any sort of test.

The test was applied to only one sentence of Russian chemical text. The sentence was deliberately chosen because it exhibited a number of difficulties which could be solved by the General Analysis Technique; these were such difficulties as reflexive verbs, inversion of subject and predicate, the interpolation of case relationships for numbers in numeral form, prepositional government, agreement stretches, and modification by adverbs. Since it was

known what the text would be, a number of ad hoc procedures were added to the general procedures and an acceptable translation was obtained. It was clearly realized, however, that, without the ad hoc measures, the translation would not have been acceptable. A report of this experiment is given in the Journal of the Association for Computing Machinery, Volume 6, Number 1.

On August 20, 1959, the code-matching group produced a machine translation ten sentences in length. The translation was excellent. No information was provided as to how it had been achieved. Other workers at Georgetown hazarded the guess that the procedures were almost entirely ad hoc, and information derived later from some members of the group indicated that this was the case. Miss Lukjanow was reticent in discussing her methods and did not produce a translation of either a random text or a prepared text of any greater length.

In the spring of 1959, the seminar papers, which had been appearing fairly regularly since September, 1956, and of which some fifty had already appeared, were replaced by a new series. While the seminar papers had reported on the organization of the various groups and on the first gropings toward a working method, the new papers described the analyses and programs which were rapidly coming to seem viable and definite. These papers were named Occasional Papers and more than twenty-five of them have been published since March 1959.

In June 1959, a series of tests was run. The General Analysis Technique was again tested. An 'examined' text - that is, one which has been checked against the machine dictionary so that any textwords not in the dictionary can be entered - was used first. This text was 100, 000 running words in length. A random text of 1500 running words was also used. Both texts were in the field of organic chemistry, and in the Russian language. A chemist who had no connection with Georgetown Machine Translation examined the English versions and concluded that the texts conveyed the essential information although their style was clumsy and the reading of them was time-consuming. Still it was significant that the information conveyed in a Russian text chosen at random was transferred into English. This was a marked advance.

At this point, the name 'General Analysis Technique' was changed to Georgetown Automatic Translation'.

At the same time, the sentence-by-sentence method was also tested. Many developments had extended the system to the point where another name seemed advisable and Dr. Brown now called his system the Simulated Linguistic Computer. He translated an examined text of 200, 000 running words and a random text of 10, 000 running words. The tests on the examined text proved at least as successful as the tests of the General Analysis Technique but the test on the random text was not quite so acceptable. Nonetheless, there was a transfer of information from the random French text into English, and the test of the Simulated Linguistic Computer is no less significant than the test of the Georgetown Automatic Translation.

The syntactic analysis group was not prepared to make a test at this time

The machine tests of the General Analysis Technique and of the Simulated Linguistic Computer revealed not only the possibilities of the two systems, but also their specific deficiencies and lacunae. The research workers began ad intensive program of correcting and completing the translation procedures.

These corrections and additions were to be machine tested at convenient intervals to assure that they operated well themselves and that they did not disturb any of the established programs.

At the end of the fiscal year in June 1959, the Georgetown Machine Translation Research Project reviewed the code-matching method. While it seemed certain that code-matching must be in any case a necessary part of machine translation, any attempt to rely only on code-matching seemed unsophisticated in view of the unavoidable clumsiness of the system of coding. The attempt to translate by a single word-by-word pass from left to right through the text seemed to involve unnecessary restriction. After the review, Georgetown repudiated the code-matching technique as a total solution of the problems of machine translation.

In July 1959, a UNESCO Conference on Machine Translation was held in Paris. Georgetown sent Mr. Zarechnak and Dr. Brown as observers. Mr. Zarechnak made a statement from the floor in which he described briefly his GAT translation of June 1959. Dr. Brown was asked, by some of the participants who were interested, to give a demonstration of the Simulated Linguistic Computer with a random French text, and he did so. This was the first public demonstration of the translation of a purely random text.

The International Conference for Standards on a Common Language for Machine Searching and Translation was held in Cleveland in September of 1959. Dr. M. Pacak attended and read a paper on: Morphology in Terms of Machine Translation. The proceedings were published in Information Retrieval and Machine Translation, Volume III, Part 2; Allen Kent, Editor; Interscience Publishers, New York, 1960.

Demonstration runs in the machine translation of organic chemistry were made at the Pentagon on January 25, 1960. The GAT process was used. The test was attended by representatives of government agencies and of various 3 institutions in the Washington area. Three demonstration runs were made.

1. A rerun of about 1, 100 words of the random text of June 1959;
2. A rerun of approximately 3, 000 words from the first corpus of 100, 000 words; and
3. A rerun of approximately 4, 500 words from the second corpus of 100, 000 words.

These runs were of the same quality as those made in June 1959, since the programs were the same in both cases.

In February 1960, A National Symposium on Machine Translation was held at the University of California at Los Angeles. Professor Dostert participated as a Chairman of one of the sessions and as a "discussant". Dr. Brown and Mr. Zarechnak described the research at Georgetown and each read a paper. Mr. Zarechnak spoke on "Nesting within the Prepositional Structure", Dr. Brown on "Flexibility Versus Speed". The gist of these papers is explicit or implicit in the descriptions of the GAT and of the SLC which are given in this book. Mr. Peter Toma, who directed the programming of the GAT (a system known as the Serna System), was also present, and participated as a "discussant".

Dr. Garvin and Georgetown University severed their connection in March 1960. The syntactic analysis method had not yet been tested. When an account of it was published subsequently, it was observable that Dr. Garvin had found it necessary to expand it beyond the limits of the original theory of syntactic analysis.

At Princeton University, in the month of July 1960, a conference of federally-sponsored machine translation workers was held. The purpose was the interchange of specific information among a small number of people working in the field. Participation was restricted to two members only from each group. Dr. Brown and Mr. Zarechnak attended as the representatives of Georgetown. Various aspects of the work were discussed in the hope that enough similarities might be found to allow of greater co-ordination among the groups. One result of the conference was the formation of a committee under the Chairmanship of Dr. H. Josselson of Wayne State University to discuss the practicability of a general agreement on the format of dictionary information and the possibility of establishing an exchange procedure. The meeting was subsequently planned for April 1961, and was to be held at Georgetown University.

As a result of discussions with Professor Dostert during his trip to Belgrade to discuss the language-teaching laboratories there, the government of Yugoslavia decided to undertake a project for machine translation from English into Serbo-Croatian. Georgetown University agreed to grant fellowships to a number of Yugoslav candidates for the degree of Master of Arts so that they might study English, Linguistics, and Machine Translation at the University. The first such students arrived for the Fall Semester.

In the last months of 1960, a Keypunch Center was organized in Frankfurt, Germany, in order to take advantage of the lower costs which might be expected in such a location. By February 1961, the organization of the Center was complete, and production began.

Very early in 1961, a considerable dislocation was caused by the need to convert the programs from use on a 705 computer to use on a 709. The

Georgetown Automatic Translation System is essentially a specially coded dictionary and a set of rules for transferring Russian structures into English structures. When the system is used in connection with a computer, naturally, computer programs are employed. The programming system for use on the 705 was named the Serna System, and was developed largely by Peter Toma. Although the translation system itself is essentially independent of machine considerations the programming system is not. Consequently, the change from the 705 to the 709 computer involved a change of all programs. Though it is theoretically possible to make such a change in a mechanical way, the possibility of increasing the accuracy of the programs and of effecting economies in the correlation of the various sections prompted a decision to rewrite all of the programs from the beginning. This process required longer than was originally estimated, and consumed over a year of time. After the programs had been rewritten, they were sufficiently different in nature that they were given a new name; they ceased to be referred to as the Serna System and were now called the Direct Conversion programming.

The SLC was also adapted at this time from use with a 704 computer use with a 709. Though the essential programming was not affected to the same as in the Direct Conversion programming, this change also required much time and energy.

In early April 1961, the working conference on Russian to English grammar codes was held at Georgetown. The participants discussed the type of dictionary codings which were in use in each of the eight centers represented, the kind of information which must be covered by a grammar code, the possible format for such a code, the organization of the coding operation, and the possibility for interconvertibility of the various coding systems. The general impression was that the eight groups worked in such different ways that interconvertibility was not readily possible. The committee in charge arranged for the circulation of questionnaires on the various coding systems, with the intention of sifting then at leisure and publishing a report.

In September 1961, an International Conference on Machine Translation was held at Teddington, near London. Professor Dostert, Dr. Brown and Mr. Zarechnak attended. Dr. Martin Joos, who is associated with Georgetown Machine Translation Research in an advisory capacity, was also present. Mr. Zarechnak read a paper: A Fourth Level of Linguistic Analysis; the contents of this paper are given later in this report as two separate papers: Syntagmatic Analysis, Noun-Noun Structures, and: Lexical Choice, Further Research Based on Order Categories of Nouns. Dr. Lawrence Summers, a chemist from the University of North Dakota, who was spending a year at the Georgetown Machine Translation Project to acquaint himself with the problems of machine translation, also read a paper on the mechanical synthesis of chemical terms. (See the paper on Morphology; Base Analysis.)

For some time now, a research group had been working on the problem of machine translation from Chinese to English. The problems involved in

the ideographic writing, in the lack of inflection, and in the sentence structure proved so different from the problems of translating from Russian or French into English that it was difficult to decide how to begin. In September 1961, Dr. John de Francis was appointed as a Consultant to the researchers in this field and began weekly discussions of the problem of analyzing Chinese from the point of view of machine translation.

Limited research in other languages had also been carried out from time to time, not with the idea of perfecting a translation involving those languages, but with the idea of investigating certain points of structure to see whether new points of view on the basic research languages could be achieved. Beginning in September 1961, a comparison of French and English was conducted to determine if French could shed any light on the patterns of distribution of the English article. A similar comparison between Turkish and English was also conducted. These languages were chosen because French, like English, has an extensive system of articles, though their patterning is different, while Turkish has no articles as such, though it does have structural categories which convey similar effects. In addition, it was felt that the work in French could constitute a development of the French-to-English work done by Dr. Brown with the SLC, and the work with Turkish would undoubtedly have some bearing on the work on machine translation which was beginning to develop in Turkey.

From the earliest days of machine translation research at Georgetown there had been discussion of the possibility of developing a system for translating not only from one language to another, but also for translating from any of a number of languages to any other. In October 1961, recent additions to the staff at Georgetown provided the personnel needed for this type of research in the Slavic Languages. The Comparative Slavic Research Group, under the direction of Milos Pacak, began by making a comparative study of Czech, Polish, Russian and Serbo-Croatian. A system of transcription compatible with the orthographies of these languages was evolved, and work was begun on the morphology of the individual languages. A number of individual papers have been published, and it is expected that the entire group may be published shortly.

The importance of this work is great. The research is not only an investigation of various Slavic Languages, but also an investigation into the possibilities of core languages for machine translation. Since the Slavic Languages selected are reasonably similar, though they belong to each of the three branches of Slavic, the possibility of establishing an artificial core language through which to transfer from one to the other and from any to English is potentially very great. It is expected that this research will be most informative on the subject of core languages.

In the latter part of 1961, the SLC programming for French-to-English translation was adapted also for Russian-to-English translation. By the beginning of 1962, the two versions of the SLC had been co-ordinated and reworked so that the SLC was no longer a system of programs for a specific translation project, but became a generalized programming language which can be adapted to any machine translation situation.

After the conference at Teddington, Georgetown had been invited by Euratom to demonstrate the Georgetown Automatic Translation System. The demonstration translation was run in December 1961. This translation made use of the SLC only, since the conversion of the Serna System (705 computer) to the Direct Conversion System (7090 computer) had not been satisfactorily completed.

The result of this demonstration was an agreement between Georgetown University and Euratom. By this agreement, Euratom was to use the GAT-SLC systems for making translations and as the basis for further research; Georgetown was to be afforded the use of the Euratom Computer at Ispra in Italy for testing and translating.

An International Symposium on Symbolic Languages in Data Processing was held in Rome in March 1962. Mr. Zarechnak attended and read a paper on: Some Operational Solutions for Multiple Meaning in Machine Translation; this paper was the product of a collaboration between Michael Zarechnak and Milos Pacak.

In April 1962, another Russian-to-English test translation was made, again by means of the SLC method of programming. The text was in the field of economics.

This translation showed a noticeable improvement over former translations, since many inadequacies which had formerly been evident disappeared. In some cases this disappearance was the result of altering the linguistic statement or the dictionary listing. In other cases, the disappearance came about automatically, once anterior errors in the system had been corrected.

During the latter part of 1961, the Joint United States Military Mission for Aid to Turkey had manifested an interest in machine translation. In May Dr. Macdonald visited Ankara and oriented and co-ordinated a staff selected for the elaboration of a pilot project for English-to-Turkish translation.

The Reverend Roberto Busa, S. J. , who is widely known for his automated lexical research on the Dead Sea Scrolls, on the Writings of Saint Thomas Aquinas and on Biblical Concordances, visited Georgetown University in May 1962, and held two seminars. The subject which Father Busa discussed was: Automation of Lexical Research; this included pre-editing of texts, keypunch formats, machine operation techniques, lemmatizing techniques, and various graphico-semantic techniques including homographic forms. The seminar meetings were open to the public, and included periods for general discussion.

In June 1962, the Chinese section published a Telegraphic-Code Chinese English Dictionary for Machine Translation. This work, compiled under the direction of Mrs. Anna Chennault, gives the telegraphic code number of each of 9699 characters, including the newer and simpler forms recently introduced by the government of the People's Republic of China; the dictionary also gives the system of Romanization recently introduced by that government.

Once this dictionary was completed, it became possible to proceed with a pilot project in the translation of Chinese to English; a text in mathematics was selected for translation, and the research was begun.

A second conference on machine translation was held at Princeton University in June 1962. The general topic for the discussion was: Syntax. Dr. Pacak and Prof. Zarechnak attended as representatives of Georgetown University.

The NATO Conference on Automatic Translation was held in Venice in the summer of 1962. Professor Dostert was invited to participate, and read a paper during the introductory sessions; this paper reviewed the procedures and results of the Georgetown Project. Dr. Brown took part in the conference also, giving a series of lectures on the SLC programming system.

A demonstration translation was run in October 1962, using a computer at Oak Ridge, Tennessee, under the auspices of the Atomic Energy Commission. An examined text of some 45, 000 words in the field of cybernetics was translated, along with a shorter random text, also in the field of cybernetics. Because of the difference in content and in level of style, both of these texts presented considerably greater problems than the texts in chemistry, physics and other similar sciences that had been translated previously. This translation showed a definite pattern of improvement in certain areas, particularly in the subject recognition routine. As usual, the translation served as the basis for intensive studies designed to formulate still more improvements in the system. The necessary research has been planned and is now under way. Papers detailing the nature of this research will shortly be published.

The Keypunch Center in Frankfurt was closed in December 1962, because the volume of keypunching required no longer justified the maintenance of a separate facility.

The liaison established with the Euratom Center in Ispra, Italy, continues and provides Georgetown with much of the machine time that is needed for the testing of the various improvements.

Recently Euratom has announced that it proposes to develop a machine translation system for translating from Russian to French. It has suggested using the GAT analysis of Russian as the basis for establishing the transfer patterns. Georgetown has been asked to send a consultant to Brussels to aid in the adapting of the GAT to this purpose. On the basis of this co-operation, negotiations have been opened for a contractual agreement between Euratom and Georgetown.

The terminal date for this report is March 31, 1963.

THE TRANSCRIPTION

A basic feature of a Russian-to-English translation system is the transcription.

Most of the computers which are regularly available in the United States are geared to the English form of the Latin alphabet. Russian uses a Cyrillic alphabet. This poses one problem. The English Latin alphabet has twenty-six letters. The Russian Cyrillic alphabet has thirty-two letters. This poses a second problem.

The problem of input, that is, of preparing the text so that it is usable by the computer, is easily solved. The keypunch machine which is used for putting ing the information on cards has a key for each letter of the Latin alphabet, for each digit, and for each of certain symbols. These keys can be equipped with Cyrillic caps and the keypunch operator can simply read off the Cyrillic characters and punch them. The fact that there may be no particular connection between a Cyrillic alphabet cap and the standard cap that it replaces is of no importance. What is important is that each symbol in the Cyrillic alphabet be consistently represented by the same pattern of punches on the cards.

However, the problem of output, that is, of inducing the computer to print out a text in Cyrillic characters, could only be solved by the use of special and more involved adaptations of the equipment, or by training copyists to convert the patterns of punches into the Cyrillic alphabet.

A more satisfactory method of dealing with the problem of the alphabets is to use a transcription substituting a given Latin letter for each letter of the Cyrillic alphabet. The advantages of this system are that the computer can print out material in the Latin alphabet and that the transcription symbols can be so chosen that the transcription is relatively easy for those who know Russian to read.

In view of these advantages, it was decided to use a transcription into Latin characters for the GAT.

Once the principle of transcription has been accepted, there are at least two patterns that may be followed. The transcription may have exclusively one-to-one correspondences in which only one individual letter or symbol of those available on the keypunch machine will serve as the transcription of each individual letter of the Cyrillic alphabet. Or the transcription may allow for certain two-to-one correspondences, in which a combination of two of the letters or symbols available on the keypunch machine will serve as the transcription for a certain single letter of the Cyrillic alphabet.

The advantage of the one-to-one system is that it makes splitting inflected items into bases and endings easier in certain respects. It also makes the transcription easier to read in general.

The disadvantage of the one-to-one system is that since the Cyrillic alphabet has thirty-two letters and the Latin alphabet has twenty-six, certain non-alphabetic symbols are needed to represent certain letters of the Cyrillic alphabet. This is not too difficult; the ten digits and such infrequent symbols as % or $ may be used if some convention is adopted which will show specifically when the digit or symbol is not a part of the transcription, but actually represents a number or a concept.

The second system, which allows of two-to-one correspondences, has the definite advantage of using only Latin letters. In the case of a language such as Russian, in which the orthography is very close to the morphophonemic system, a transcription which allows two-to-one correspondences can be manipulated by the linguist to provide certain useful patterns for morphological analysis; however, this sort of transcription may be less advantageous than a one-to-one transcription in the more advanced stages of translation; the relative merits of each type of transcription at all points of the translation system must be carefully weighed.

The GAT uses a one-to-one transcription. Those Cyrillic letters which have a counterpart in shape or in sound in the Latin alphabet are transliterated by that counterpart. Those Cyrillic letters which have no counterpart in the Latin alphabet are transliterated either by one of the remaining letters in the Latin alphabet or by one of the digits. In order to facilitate memorization of the GAT transliteration system, these letters and digits were chosen for their similarity in form to the Cyrillic letter which they transliterate.

The fact that some of the Cyrillic letters are represented by digits will cause no confusion in the translation process; if a digit is intended to represent a number, it is preceded by a symbol which indicates that what follows is a number, while, if the digit is intended to represent a letter, either there will be no preceding symbol, or there may on occasion be a symbol to indicate that the digits which appear in the text represent the characters of the Cyrillic alphabet.

The transliteration system is as follows; only capital letters are used, since lower case letters are not available.

1.	А	A	9.	И	I	17.	Р	R	25.	Ш	W
2.	Б	B	10.	Й	1	18.	С	S	26.	Щ	5
3.	В	V	11.	К	K	19.	Т	T	27.	Ъ	7
4.	Г	G	12.	Л	L	20.	У	U	28.	Ы	Y
5.	Д	D	13.	М	M	21.	Ф	F	29.	Ь	6
6.	Е	E	1М.	Н	N	22.	Х	X	30.	Э	3
7.	Ж	J	15.	О	0	23.	Ц	Q	31.	Ю	H
8.	З	Z	16.	П	P	24.	Ч	C	32.	Я	4

For the work in Comparative Slavic, where the transcription system may cover four languages at present (Russian, Polish, Czech, Serbo- Croatian) and perhaps more later, the one-to-one transcription used for the GAT is less useful. This is due in part to the fact that the computer now being used allows for only forty-three different characters, including letters, digits, symbols and punctuation marks. Forty-three characters are not enough to represent the various letters which occur in the alphabets of these four languages, even though there is considerable overlap at points of similarity. Therefore, a new transliteration system in which two-to-one correspondences are permitted has been evolved, and is now being tested in various forms to see what advantages can be gained from such a system.

Transcription of Punctuation, Numbers and Non-Cyrillic Letters

The computer permits the use of only forty-three different characters. Since these include the twenty-six letters of the Latin alphabet (capitals only) plus the ten Arabic digits, there are only eight characters left for use as punctuation marks. These may be varied to suit the demands of the user.

In such a situation, a new system of punctuation has been evolved by combining the available characters into complex groups. An exhaustive description of all of these combinations is beyond the scope of this paper, but some specific examples will show the nature of the various problems and the way in which they have been handled.

The beginning of each new paragraph is signaled by the symbol P* as a separate item of the text.

Titles, subtitles or descriptive lines under diagrams and tables may have their beginnings marked as paragraphs and will have their endings marked as sentence terminals.

The beginning of a new page is signaled by the word PAGE in parentheses as one item, followed by the number of the page (Arabic numerals have the symbol $ suffixed) as a second item, followed by the asterisk which marks the end of a sentence as the third item. Page notations are thus sentences in then selves and are inserted before the beginning of the first complete sentence on the page in question.

All sentences have their ending signaled by an asterisk. If the sentence in the text ends with a period, the asterisk symbolizes and replaces the period If the sentence in the text ends with some other mark of punctuation, the symbol for that mark of punctuation is inserted as a separate item after the last word, and the asterisk is inserted as another separate item after the symbol for the mark of punctuation. If the sentence in the text does not end with any mark of punctuation (because it leads into a diagram, or because it is a title) the symbols N and * are inserted as two separate items after the last word, and the asterisk is inserted as a separate item after the first asterisk, so that the aggregation is N * * .

If a period occurs within a decimal number or formula or after an abbreviation, it is entered as a period. It may form a separate item or may be one with the accompanying form depending on the spacing in the text.

A comma is keypunched as a separate item. If a comma occurs as an integral part of an independent item or as a sign of abbreviation, it is keypunched so that the whole remains as one item.

The colon is punched as a separate item; two successive periods (..) are used.

The semi-colon is punched as a separate item; a period and a following comma (. ,) are used.

Quotation marks are punched as a separate item; they are distinguished as to whether they open or close the quotation; double parentheses are used (((or))).

The question mark is punched as a separate item; a virgule, a G and a period (/G.) are used.

The dash is punched as a separate item; one hyphen (-) is used. A dash in the text is marked by blank spaces preceding and following it, since it is a separate item.

A hyphen is punched with the matter to which it relates so that the whole is one item; one hyphen (-) is used. A hyphen is never preceded or followed by a blank space. A required hyphen which happens to occur at the end of a line is to be differentiated from the facultative hyphen which serves as a mark of syllable division at the end of a line in the text. It may be necessary to distinguish these for the keypunch operator during a preliminary reading of the text.

A capital letter is keypunched with the prefixed symbol C$; an item which is all in capital letters is keypunched with the prefixed symbol CC$.

A Latin letter is keypunched with a prefixed symbol L$. When a letter occurs in such a position that it is not clear whether it is a Cyrillic or a Latin letter, it is keypunched as a Latin letter.

Every Arabic numeral is keypunched with a suffixed symbol $.

Every Roman numeral is keypunched with a prefixed symbol R$.

A relatively uncomplicated formula is keypunched as it stands and has the Prefixed symbol F$; superscripts and subscripts are treated as on-the-line characters and lower case letters have a prefixed comma. A complicated chemical formula is not keypunched. Instead, the symbol FBHF (fill-by-hand formula) is keypunched and the approximate space occupied by the formula is

given as a rough precentage of the page of the text. Formulas which occupy separate lines are keypunched as separate sentences.

The keypunch system has had three forms in the course of the history of the GAT. The present form is known as System C.

For those who may be interested in the particulars, System C is given in detail in the General Information and Operation Manual for Georgetown Automatic Translation: J. A. Moyne; Georgetown University, 1962.

THE MACHINE DICTIONARY

It was decided at the beginning of the Georgetown Machine Translation project that the machine dictionary was to be text-oriented, that is, compiled entirely from primary sources such as periodicals and scientific publications in Russian and English.

Materials for compiling the dictionaries were selected at various times by the sponsoring agency, and now run to:

324, 500	running words in Organic Chemistry
750, 000	Economics
490, 000	Physical Chemistry
583, 000	High Energy Physics
150, 000	Celestial Physics
445, 000	Meteorology and Geodesy
45, 000	Cybernetics

The Original Procedure

Organic Chemistry was the first discipline processed; items abstracted from the first 35, 000 running words selected from the Soviet Union's Journal of General Chemistry form the basis of the original GAT machine dictionary. The procedure for compiling the original Dictionary was as follows:

1) The translation into English of the Journal of General Chemistry, prepared by human translators for commercial use, was revised so as to reflect the original Russian text faithfully, since the unrevised translation was so full of inconsistencies--and even errors--that it was unsuitable for dictionary abstraction.

2) The sentences of the Russian text were numbered serially; the corresponding English sentences were also numbered serially.

3) Every Russian item in the text was entered on a separate card and the cards were alphabetized.

4) The stems of the Russian split items and the full forms of the unsplit items were entered on dictionary sheets and coded.

5) One English equivalent (gloss) of each Russian entry was located in the English text and entered on the dictionary sheet. These English glosses were entered in their canonical form and were coded for English synthesis.

6) The information from the dictionary sheets was then keypunched on 80-column punch cards.

The Modified Procedure

The remainder of the Organic Chemistry material was processed for the dictionary in five different batches; with each batch, the existing dictionary was updated.

Some of the procedures for abstracting items for the dictionary were necessarily modified.

1) The computer was used for sorting, for comparing and for morphologically analyzing the Russian text. This work had formerly been done by hand.

2) The keypunched Russian text was sequentially numbered, alphabetically sorted and morphologically analyzed, and was then compared with the existing dictionary.

3) Those items which already existed as entries in the dictionary were disregarded.

4) Those items which did not already exist as entries in the dictionary were printed out in alphabetical order and with their sequence number as 'not-in-dictionary items'; these became known as 'error listings' from the fact that they were treated as 'errors' in the morphology program. Thus the language analyst, was presented with a list of the error listings and with the printout of the transliterated Russian text. Each line of the text carried the sequence number of the last item in the line. In order to locate the English equivalents of these items the following procedure was used:

    a) The pages of the transliterated Russian printout text were numbered

    b) A table was set up showing each page number accompanied by the first and the last sequence number on that page.

    c) The pages of the original Russian text were collated with the pages of the transliterated Russian text, and this information was also entered in the table.

    d) The pages of the English translation text were collated with the pages of the original Russian.

    Thus, the table finally held information of this type:

    Item (sequence number 1715) is on page No. 75 in the transliterated Russian text; this corresponds to page No. 15 in the original Russian text and to page No. 10 in the English translation text.

5) The researcher then continued with the original procedure, beginning at point 4. The table described above was used to establish the English gloss.

The Present Procedure

When the Economics Dictionary was begun, the procedure described above was discontinued because it was too slow and did not provide enough information about the range of translation of the individual Russian item.

The following procedure was introduced and is now in use.

1) The Russian text is printed out a line at a time. The printout is not consecutive; each line repeats a portion of the preceding line in such a manner that the first letter of every item in the text comes to stand at a specified point in the center of a line. Thus, when an item begins at the specified central position, its environment is given to its left and to its right. The limits of the environment are set by the width of the paper. The lines are then sorted alphabetically on the basis of the center item, and the sorting is printed out. Such a printout is called a full concordance; it gives the language analyst the possibility of studying each item in all of its environments in the text. The full concordance has proved to be a useful analytical tool in almost every phase of machine translation.

2) A concordance of error listings (error concordance) is prepared in the same manner as the full concordance.

3) The error listings are numbered in sequence and every new item in the running transliterated text has its error number printed before it.

4) The language analyst has a set of punch cards with prepunched sequence numbers in ascending order. He scans the transliterated Russian text and its English translation simultaneously. If he encounters an item preceded by an error number, he writes the English equivalent on the card bearing the same number as the Russian item.

5) This written information is then punched on the card, and the information from these cards is transferred to the error concordance. Thus, an error concordance complete with English meanings is evolved.

6) A full range of translations for each individual item is obtained in this manner. The researcher studies this range and selects the equivalent which will cover the greatest number of occurrences. Those occurrences which cannot be covered by one equivalent are subjected to further study.

A number of other improved procedures have been suggested by the dictionary workers. These procedures remain to be tested before they can be put into use. Papers describing those which prove useful will be published shortly.

The Form of the Dictionary

The dictionary may appear in a number of forms, depending on the work for which it is required. The initial form is that which is punched on cards. Since any further developments are simple mechanical rearrangements of the material on the cards, it is easiest to describe the card dictionary first, and then to state how any variant forms differ from it, when these forms become important to the discussion.

The dictionary entries are punched on cards. Corrections and additions are made at any time by adding new cards. The machine sorts these, files them, and removes those cards that have been superseded.

In order to describe the dictionary, it is necessary to observe two pail of distinctions. There is the important distinction between the split and unsplit dictionaries. There is the less important distinction between the hand-entered, or permanent codes, and the machine-generated, or temporary cod

The Split and Unsplit Dictionaries

The overall dictionary is divided into two chief parts. These are the split and unsplit dictionaries.

The Unsplit Dictionary

The unsplit dictionary contains two classes of items; there are inflected and uninflected forms. The uninflected forms are adverbs, conjunctions, particles, prepositions, certain classes of numerals and pronouns, abbreviations, the numeral form of the numbers through ninety-nine, chemical formulae, words in Latin letters, and the Greek letters. The inflected forms are:

a) items which participate in the exclusion and idiom routines;

b) those items which do not participate in the exclusion and idiom routines, but which, if entered in the split dictionary, would be analyzed as if they participate in them;

c) full forms of words consisting of a variant stem with a standard ending, where the variant stem is formed by mutation (M5- from MST-) or by vowel insertion (ZEMEL6 from ZEML-), and where the distribution of the variant stem is not wide;

d) past tense masculine singular verb forms which have a zero suffix (MOG);

e) all forms of an item whose declension or conjugation is so erratic that inclusion of the item in the split dictionary would be uneconomical (4, MEN4, MNE; DAM, DAW6, DAST).

All these items, whether inflected or uninflected, are entered in the dictionary severally and at full length, and this is the reason that the dictionary is called the unsplit dictionary.

The Split Dictionary

The split dictionary contains those items which are inflected and whose inflection pattern is sufficiently regular that it is more economical to store only their stems in the dictionary and to list their endings in a separate dictionary of endings. All items in this dictionary are entered as stems, the inflectional endings having been removed, and this is the reason that the dictionary is called the split dictionary. The stem is the longest part of a word common to all of its inflections, as CITA- in CITAT6, CITAET, CITALI. Sometimes it is more economical to enter two stems, as BRA- and BER- in BRAT6, BERET, BRAL1; a certain number of verbs are entered with three, four or five stems, though these are by no means numerous; an example with three stems is MO-, MOJ-, MOG-, as in MOC6, MOJET, MOGLI.

The distinction between the split and the unsplit dictionary is most important in the dictionary lookup routine. After the lookup, all items are treated in exactly the same way. In coding the dictionary, very few distinctions are made between the coding of a split entry and the coding of an unsplit entry. Where such distinction are important, they usually involve a difference between codes which are hand-entered or permanent, and codes which are machine-generated, or temporary.

Permanent and Temporary Codes

Codings are either permanent or temporary. Permanent codes represent information which is always true in respect to the item which receives the code. These codes are an essential part of the dictionary. They are hand-entered by being written on dictionary format sheets and then keypunched on the cards. Temporary codes represent information which applies to the item only in a specific environment. These codes do not appear in the dictionary. They are generated by the computer during the transfer process and they are put into certain positions in the work space of the material which has been copied from the dictionary. They complete the analysis of the specific occurrence of a word so that it can be transferred and translated in reference to that particular environment. In the description of the dictionary which follows, the distinction between a position which receives a permanent code and a position which is

reserved for a temporary code is noted where necessary.

The Basic Form of the Dictionary

The dictionary is keypunched on IBM cards. There are eighty available positions on each card. A dictionary entry requires three cards if it is a split dictionary entry, and four cards if it is an unsplit dictionary entry.

For convenience of reference, a numbering system is used. The cards for each entry are assigned Roman numerals. Since there are no more than four cards, only the Roman numerals from one to four are required. The positions on each card are assigned Arabic numerals. Since there are eightly positions on each card, the Arabic numerals from one to eighty are required. A colon is used to separate the Roman numeral from the Arabic numeral. Thus, I:1 denotes the first position on the first card; III:80 denotes the eightieth, or last position on the third card; II:7-12 indicates all positions from the seventh to the twelfth inclusive on the second card.

(In earlier publications, this notation employed only Arabic numerals and hyphens. The three references given above were written as 1-1, 3-80, and 2-7--12 in the old system.)

The dictionary entries are entered on the cards in the following manner.

The First Card.

Positions I:1-33 receive the Russian entry. The Russian entry is left justified, that is, it is so placed that it begins in the first position on the left (in this **case,** in position I:1), and extends as far to the right as necessary. Any unused positions are left blank.

Position I:34 receives the code N if the Russian entry is split, and if there is a following split entry with the same stem, but with different endings. This **code** indicates that the lookup of the first entry is not to be final if there has been a matching, but that the succeeding entry must be scanned also. Otherwise, **this** position is blank.

Position I:35 receives the code 1 to indicate that this is the first card of the three or four which comprise the dictionary listing.

Position I:36 receives the parcue. The parcue indicates the part of speech **of** the Russian entry.

0	non-Cyrillic item		
1	noun	5	preposition
2	verb	6	conjunction
3	adjective	7	particle
4	adverb		:

(Later, in the dictionary lookup, if a word is not found in the dictionary, the computer will generate the code 8 in the parcue position to signify that the word is not in the dictionary. This has nothing to do with the permanent coding of the dictionary.)

Positions I:37-44 receive a variety of codes. The most important is the parset code.

The parset (an acronym for 'paradigmatic set') indicates the inflection of the Russian entry. The parset must be read in connection with the parcue, since the value of a coding in any position varies depending on whether it is applied to a verb, a noun, an adjective or some other part of speech.

0) If the entry is a non-Cyrillic item (Parcue 0), positions I:37-39 receive a three-digit interpolation code which is explained in the paper on Interpolation.

1) If the entry is a noun (Parcue 1), position I:37 receives a code indicating gender (0, any gender; 1, masculine; 2, feminine; 3, neuter). Position I:38 receives a code indicating animateness (1, animate; *2,* inanimate). Positions I:39-44 receive codes indicating the declensional subclass to which the entry belongs; these codes are discussed in the paper on Morphology.

2) If the entry is a verb (Parcue 2), position I:37 receives a code indicating the form of the verb (1, infinitive; 3, participle; 4, gerund; 6, finite form). In the unsplit dictionary, any one of the four codes is entered. In the split dictionary, only codes 3 and 6 are entered with 6 representing all forms which are not participles. (Later, in the course of the morphological analysis routine, the computer will generate the code 1 or the code 4 if it analyzes a form as an infinitive or as a gerund, and will replace the code 6 in its working material by the code 1 or the code 4, whichever is appropriate. But this has nothing to do with the permanent coding of the split dictionary.)

Position I:38 receives a code indicating the nature of the participle (1, past passive; 2, past active; 3, present passive; 4, present active) or a code indicating possible ambiguity between the short form masculine of the present passive participle and the first person plural of the non-past tense (A).

Positions I:39-44 are not needed for further parset codings for the Russian verb and are free to receive codes indicating the inflection of the English verbs in the English meanings. Position I:39 receives a code which indicates the inflection of the verb in the second English meaning in cases where there is lexical choice. See the paper on Transfer and Synthesis.

Position I:40 receives a code indicating how the reflexive suffix (-S4, -S6) is to be transferred (1, to be transferred by the English passive; 2, to be transferred by the English active; 5, to be transferred with the second English meaning,

which is usually an adjective in this case. Codes 3 and 4 have fallen into disuse (See the paper on Transfer and Synthesis).

Positions I:41-44 receive a four-position English verb synthesis code which indicates the formation of the various forms of the English verb in the first English meaning listed. (See the paper on Transfer and Synthesis.)

3) If the entry is an adjective (Parcue 3), position I:37 receives a code indicating whether the adjective does (code 1) or does not (code 0) take a comparative inflection.

Position I:38 receives a code indicating whether there is a possible ambiguity between the comparative and the long form neuter nominative and accusative of the positive (code 1), or not (code 0). Positions I:39-44 receive codes indicating the declensional subclass to which the entry belongs; these codes are discussed in the paper on Morphology.

4-7) Adverbs (Parcue 4), and all parts of speech with parcues larger than 4, have no inflections or subdivisions which are distinguished by parset codes. With these parts of speech, therefore, positions I:37-44 remain blank. Entries in the unsplit dictionary do not undergo morphological analysis, and the parset code of any part of speech in the unsplit dictionary does not need to be as long; thus, only a portion of this field is used even in the case of nouns and adjectives.

Positions I:45-46 are left blank. (Later, during the Dictionary Lookup, the computer will generate, in its working material, codes which indicate the | length of the Russian entry.)

Positions I:47-48 receive Exclusion Candidate and Boundary codes. (Since the exclusion routine deals only with full forms, these codes are found only in the unsplit dictionary.)

Position I:47 receives a single-character code to indicate exclusion boundary possibilities.

A Entry may be right or left boundary.
L Entry may be left boundary.
R Entry may be right boundary.

Position I:48 is keypunched with X if the Russian entry is a candidate for the exclusion routine. (See the paper on Exclusion.)

(Later, during the exclusion routine, if the computer analyzes the text-word as part of an exclusion, it will generate the code E in its working material in the position corresponding to I:48. But this has nothing to do with the permanent coding of the dictionary.)

Since relatively few entries receive exclusion and boundary codes, positions I:47-48 also receive, alternatively, codes which are used in the lexical choice routine. These identify the entry so that its presence in the environment can be detected more easily by searching out this two-letter code, than by the more tedious project of searching out the variable-length word itself. Words having such a code may not be subject to lexical choice themselves, but they are likely to determine an environment which will favor one lexical choice over another in some other word. Such codes, known as environment-determining codes, are incompatible with exclusion codes and there can be no conflict if they are placed in these positions. (See the paper on Lexical Choice.)

Position I:49 receives the Idcue or Idiom Candidate code. The code 1 is key-punched here if the entry can take part in an idiom. (See the paper on Idioms.)

Positions I:50-53 are keypunched with the idiom code for initials.

Positions I:54-57 are keypunched with the idiom code for sequents. If there is an idiom code for initials in I:50-53, position I:57 will receive either the code X, which indicates that no adjective may be inserted into the idiom, or the code Y, which indicates that an adjective may be inserted into the idiom. (See the paper on Idioms.)

Positions I:58-62 receive strong case determiner codes.

Strong case determiners are those forms of Russian which require, as a bound dependent, a nominal structure in a certain case. This relationship is known as strong government. Position I:58 receives the code 2 if the Russian entry is a strong case determiner of the genitive case. Position I:59 receives the code 3 if the Russian entry is a case determiner of the dative case. The pattern is repeated for the other three cases.

Position I:60	4	accusative
Position I:61	5	instrumental
Position I:62	6	locative

If a form determines more than one case, this information can be entered easily, since each case is coded in a different position.

Positions I:63-67 receive weak case determiner codes.

Weak case determiners are those forms of Russian which must be followed by a preposition governing a nominal construction. The coding of the case governed by the preposition is parallel with that for strong case determiners.

Position I:68 receives a numerical code which indicates whether a verb is transitive and what type of object it takes if it is transitive. Transitive verbs which govern the accusative (when affirmative) receive the code 1. Intransitive verbs receive the code 2. (If a verb has a reflexive suffix, in the course of the dictionary lookup, the machine will generate the code 3 in its working material

to indicate this fact, and the machine-generated code replaces any keypunch code which previously occupied this position. But 3 is never coded in the system dictionary.) Russian verbs whose English gloss is a verb which is followed by another verb without the infinitive marker 'to' receive the code 4. Russian] verbs which govern the instrumental case have the code 5 keypunched here.

Position I:69 is reserved for semantic codes where the semantic reference is to space. These codes have not yet been developed.

Position I:70 is reserved for semantic codes where the semantic reference is to measurable referents. These codes have not yet been developed. This position also receives codes 1-6 which are used in the transfer of the instrumental case ending.

Position I:71 is reserved for semantic codes where the semantic reference is to time. These codes have not yet been developed. This position also receives the Sentence Separator Codes. (See the paper on Syntax.)

Position I:72 receives a code which indicates the formation of the plural in the English gloss, if it is a noun, or the formation of the comparative and superlative in the English gloss, if it is an adjective. (This location is named Noun or AdjR, depending on the circumstances.)

Position I:73 receives the code X if the English gloss contains the verb 'to be' or receives the code N if the English entry requires the indefinite article in: the form 'an' rather than in the form 'a'. The X and N codes are mutually exclusive, and there can be no conflict. (This position is named BeLoc.)

Positions I: 74-78 receive a series of codes used by the Lexical Choice Routine and known as the Lexical Choice Routine numbers. (See the paper on Lexical Choice.)

Position I: 79 receives a code which governs the insertion of the article in English. (See the paper on Article Insertion.)

Position I:80 receives a code which governs the rearrangement of the English entries into more acceptable order. (See the paper on Rearrangement.)

The first dictionary card is now full.

The Second Card

The second dictionary card has, in positions 1-34, exactly the same information as is found in I: 1-34.

Position II:35 receives the numeral 2 to indicate the second of the dictionary cards.

Positions II:36-44 receive no codes. (Later, in a number of different routines, this space will be needed for the insertion of English words. This is done by the computer, and has nothing to do with the coding of the dictionary. The inserted words appear in various types of printout, however. If the inserted word is a preposition, its position is named EPREP (English preposition). If it is an article, its position is named DINDA (definite and indefinite article). Of course, EPREP and DINDA may co-occur in that order.)

Positions II:45-80 receive the first English meaning. The English meaning is right justified, that is, it is so placed that the last letter will be in the last position on the right, in this case, II:80. Any unused positions are left blank; the unused positions may be utilized later for additional insertions.

The second dictionary card is now full.

The Third Card

The third dictionary card has, in positions 1-34, exactly the same information as is found in I:1-34.

Position III:35 receives the numeral 3 to indicate the third of the dictionary cards.

Positions III:36-40 receive no codes. (Later, in the synthesis routine, this space will be used for adding endings to the English meaning, or otherwise modifying it. This is done by the computer and has nothing to do with the permanent coding of the dictionary.)

Positions III:41-50 receive the overflow of the English meaning. It has already been pointed out that the First English Meaning is right justified so that the point of synthesis -- the point at which inflections are to be added or changed -- is always in a predictable position (II:80). If the English meaning consists of two or more words of which some word other than the last is subject to inflection, the word which is subject to inflection terminates in II:80, and the succeeding words are entered as overflow. For example, if the English translations 'turn out' and 'kick the bucket' are used, 'turn' and 'kick' are keypunched in II:77-80, where they are available for the addition of personal or tense endings, while 'out[1] and 'the bucket' are keypunched in III: 41-50, left justified.

Positions III:51-64 receive the second English meaning, where a second meaning is necessary. The meaning is keypunched and is right justified so that it can be inflected. There is no overflow area for the second meaning. If more than two meanings prove necessary, the excess over two is stored in subsidiary dictionaries. (See the paper on Lexical Choice.)

Position III:65 receives a record mark which is a function of the computer and which has nothing to do with machine translation.

There is no further permanent coding in the split dictionary. All of the information described below is derived by the computer during the morphological analysis of the split entries, and the computer generates the necessary code for these positions.

But the entries in the unsplit dictionary are not subject to morphological analysis and so all morphological information must be permanently coded for each unsplit entry.

Positions III:66-71 receive codes indicating singular case form.

> A combination of digits indicates the case or cases of the singular which a Russian noun-form or adjective-form represents. The code 1 in the first position indicates nominative, the code 2 in the second position indicates genitive, and so on for the remaining four cases. Thus, a form such as MNE, which can be either dative or locative (preposition) singular, is coded, 003006, and a form such as TAKSI, which can be any of the singular, is coded as 123456.

Positions III:72-77 receive codes indicating plural case forms. The codes are the same as these in the immediately preceding section. The difference is, as the name indicates, that here the codes show which cases of the plural the *m* Russian entry represents. The code for TAKSI is again 123456; the code for DEVUWEK is 020400; the code for DVUX is 020406.

Position III:78 receives a code to indicate whether the form is singular (1),| plural (2), or either (3).

Position III:79 receives a code to indicate whether the form is masculine (1), feminine (2), or neuter (4). If the textword may be of more than one gender, the appropriate combination code may be arrived at by adding the single code.

Position III:80 receives a code to indicate whether an adjective is in the long (attributive) form (1) or the short (predicative) form (2).

The third dictionary card is now full.

The Fourth Card

The fourth dictionary card has, in positions 1-34, exactly the same information as found in I: 1-34.

Position IV: 35 receives the numeral 4 to indicate the fourth dictionary card.

Position IV: 36 receives a code to indicate whether an adjective is animate (1), inanimate (2), or either (3).

Position IV: 37 receives a code to indicate whether a verb is passive (1) or active (2).

Position IV: 38 receives a code to indicate person if the entry is a verb: first person (1), second person (2), third person (3), all persons (6).

Position IV: 39 receives a code to indicate the tense if the entry is a verb: past (1), non-past (2), future (with BUDU, BUDET, etc. only) (3).

Position IV:40 receives a code to indicate the mood, specifically the imperative (7), if the entry is a verb, or to indicate the comparative form if the entry is an adjective (C).

Positions IV:41-80 receive no codes.

The fourth dictionary card is now full.

Modifications

For the actual translation process, the dictionary is copied onto magnetic tapes. At least two tapes are necessary, since the split dictionary must be separate from the unsplit dictionary.

These tape recordings are the dictionary which the computer uses during the lookup process. Whenever the number of changes warrant it, the card dictionary is re-recorded, and new tapes are made.

When the dictionary is copied onto tape from the cards there is a minor repositioning. The repetitions of the Russian words and the card number codes found at the beginning of the second, third, and fourth cards are removed and the gaps closed up. Except for this, the relative spacing is exactly the same as on the cards.

For the reference of the staff who work on the dictionary, printout sheets are made from the taped dictionary. These are bound into books for convenience, and they provide a record of the state of the card dictionary at the time of the latest taping. A printout of the dictionary from the tape preserves the spacing of the tape, however, not that of the cards, and this must be taken into consideration when reading the printout.

During the dictionary lookup, as each textword is matched by a dictionary entry, the entire dictionary listing, except, of course, the Russian entry, is copied from the dictionary tape and placed after the textword. The same spacing is kept. Free space is available after the copy of the dictionary listing. The textword, the copy of the dictionary listing and the free space constitute the working material of the computer. As the computer runs through the routines, it searches out the information copied from the dictionary listings, and it generates codes. These generated codes sometimes replace codes originally

keypunched in the dictionary, but more usually they occupy some of the free space. At various stages of the translation process, all of this material can be printed out to show how that stage of the translation was reached. The *a* of this printout may vary widely as to spacing depending on whether the 705 or 709 computer was used, but the order is the same as in the original dictionary listing, and, though stretched or compressed, is easily recognizable.

Although reference numbers such as I:36 are strictly applicable only to the card form of the dictionary, they are sometimes used loosely to indicate corresponding positions in the various copies or printouts even though the 1 spatial relationships have been altered in transit from one form to another. This looseness actually conduces to clarity in most cases. However, the reader should remain aware that various formats are being discussed.

Where specific dictionary positions have been assigned names which describe the type of information coded in that position, the names have been given in the above description. See positions I:72, I:73, and especially II: 36.

The Present Status of the Dictionary

The present GAT dictionary has 37, 933 entries. This total is divided disciplines as follows.

Chemistry	10,287
Economics	9,500
Physical Chemistry	4, 288
High-Energy Physics	3,001
Nuclear Physics	5 507
Celestial Physics	1,250
Meteorology and Geodesy	3,250
Cybernetics	850
TOTAL	37,933

Disciplines

Discipline: CHEMISTRY
Running words: 324, 500
Abstracted words: 12, 231
New words: 10, 287
Sources:

1) JURNAL ORGANICESKO1 XIMII TOM XXII; Vol. 9(1952)
2) POROWKOVA4 METALLURGI4
3) KATALICESKOE ALKILENIROVANIE FENOLA
4) MINERALOGICESKIE RAZNOVIDNOSTI KVARQA
5) TEORETICESKIE PROBLEMY ORGANICESKO1 XIMII
6) PUTOXIN: ORGANICESKA4 XIMI4
7) SOSTO4NIE TEORII XIMICESKOGO STROENI4
8) TERENTOEV: NOMENKLATURA ORGANICESKIX SOEDINENI1
9) JURNAL ORGANICESKO1 XIMII TOM XXVIII; Vol. 9 (1958)

10) JURNAL ORGANICESKO1 XIMII TOM XXVIII; Vol. 11 (1958)
11) JURNAL ORGANICESKO1 XIMII TOM XXVIII; Vol. 9 (1958)
12) JURNAL ORGANICESKO1 XIMII TOM XXIX; Vol. 3 (1958)

Discipline: ECONOMICS (April 3 - June 10, 1961)
Running words: 750,000
Abstracted words: 106, 396
New words: 9, 500
Sources:

1) POLITICESKA4 3KONOMI4 pp. 64 - 555
2) VOPROSY 3KONOMIKI No. 1 (1958)
3) VOPROSY 3KONOMIKI No. 2 (1958)
4) VOPROSY 3KONOMIKI No. 3 (1958)
5) VOPROSY 3KONOMIKI No. 4 (1958)
6) VOPROSY 3KONOMIKI No. 5 (1958)
7) VOPROSY 3KONOMIK1 No. 10 (1958)
8) VOPROSY 3KONOMIKI No. 11 (1958)
9) VOPROSY 3KONOMIKI No. 12 (1957)

Discipline: PHYSICAL CHEMISTRY (Sept. 19 - Nov. 9, 1961)
Running words: 583,000
Abstracted words: 36, 005
New words: 3, 001
Sources:

1) USKORITELI 3LEMENTARNYX CASTIQ
2) LIVINGSTON: USKORITELI
3) BATLER: 4DERNYE REAKQII SLIVA
4) KAULING: MAGNITNA4 GIDRODINAMIKA
5) DORMAN: VARIAQJ.1 KOSMICESKIX LUCE1
6) PROBLEMY SOVREMENNO1 FIZIKI Vol. 1 (1954)
7) PROBLEMY SOVREMENNO1 FIZIKI Vol. 7 (1954)
8) PROBLEMY SOVREMENNO1 FIZIKI Vol. 5 (1954)
9) PROBLEMY SOVREMENNO1 FIZIKI Vol. 4 (1954)

Discipline: NUCLEAR PHYSICS (Nov. 29 - Dec. 15, 1961)
Running words: 535,000
Abstracted words: 69,732
New words: 5, 507
Sources:

1) PR1BORY I TEXNIKA 3KSPERIMENTA
2) PRIBORY I TEXNIKA 3KSPERIMENTA
3) PRIBORY I TEXNIKA 3KSPERIMENTA

4) JURNAL 3KSPERIMENTAL6NO1 I TEORETICESKO1 FIZIKI; TOM 38, Vol. 4(1960)
5) JURNAL 3KSPERIMENTAL6NO1 I TEORETICESKO1 FIZIKI; TOM 38, Vol. 5 (1960)
6) JURNAL 3KSPERIMENTAL6N01 I TEORETICESKO1 FIZIKI; TOM 38, Vol. 6 (1960)
7) JURNAL 3KSPERIMENTAL6NO1 I TEORETICESKO1 FIZIKI; TOM 39, Vol. 1 (7) (1960)
8) JURNAL 3KSPERIMENTAL6NO1 I TEORETICESKO1 FIZIKI; TOM 39, Vol. 2 (8) (1960)
9) JURNAL 3KSPERIMENTAL6NO1 I TEORETICESKO1 FIZIKI; TOM 39, Vol. 4 (10) (1960)
10) ATOMNA4 3NERGI4, MART, TOM 8, Vol. 3(1960)

Discipline: CELESTIAL PHYSICS (Dec. 13 -Dec. 22, 1961)
Running words: 150,000
Abstracted words: 10,819
New words: 1,250
Sources:

1) B. LOVELL: METEORNA4 ASTRONOMI4 (1958)
2) VASILIEV I PRESSMAN: METEOROLOGI4 I ATOMNA4 3NERGI4 (1959)
3) RIL I DR.: STRO1NOE TECENIE (1959) I

Discipline: METEOROLOGY AND GEODESY (Dec. 19, 1961 - Jan. 5, 1962)
Running words: 445, 000
Abstracted words: 39,488
New words: 3, 250
Sources:

1) DUBINSKY I DR: METEOROLOGI4 (1959)
2) LANDIN: OSNOVY DINAMICESKO1 METEOROLOGII (1955)
3) TRUDY QENTRAL6NOGO NAUCNO-ISSLEDOVATEL6NOGO INSTITUTA GEODEZII: Vol. 121 (1957)
4) LUKAVCENKO: GRAVIMETRICESKA4 RAZVEDKA NA NEFT6 I GAZ; pp. 13-11, 57-204

Discipline: CYBERNETICS (April 2 - April 13, 1962)
Running words: 45,500
Abstracted words: 3,673
New words: 850
Sources:

1) ROVENSKll I DR.: MAWINA I MYSL6

THE DICTIONARY LOOKUP

Method

In human translation, dictionary lookup is effected by the translator. Because of his experience, he does not need to look up all of the words in the text; he knows many of them already. He looks up only the words he wishes to check. He usually looks them up in the order in which he meets them in the text. When he resorts to the dictionary, he is quickly able to locate a specific item by skipping about and by making reliable estimates as to how great each successive skip needs to be. For a human being, a system of text-order lookup is efficient.

In machine translation, dictionary lookup is effected by the computer. The computer begins each translation with no memory and no experience; it must check every occurrence of every word in the text. By and large, the computer cannot skip about in its dictionary, but must check every item in order until the desired item is found. {It is possible to arrange for specific points in the machine dictionary to be so marked that the computer can make certain fixed skips in much the same way as a book dictionary can be equipped with a thumb index for the human translator. Machine "memories" which permit such skipping are not in common use; they are expensive.) For a machine then, a system of text-order lookup is inefficient; it requires too much time, or, if provisions are made to shorten the time, it becomes too expensive.

At present, the most efficient system for dictionary lookup by computer is lookup by sorting.

Each item of the text (textword) is assigned a number which indicates its serial position in the text. If a word occurs several times in the text, it is several textwords, each with its own serial number. Then the textwords are sorted into alphabetical order. This brings all of the textwords into an order which parallels the order of the dictionary entires. The sorted textwords are compared with the dictionary entries by progressing in one pass from the beginning to the end of the dictionary with only minor skips backward. When the lookup is complete, the textwords, each with its pertinent information as garnered from the dictionary, are re-sorted into numerical order on the basis of the serial numbers. The text is thus restored to its original order, but each word is now accompanied by the dictionary information.

The GAT Lookup

The original GAT lookup routine (Serna System), established by P. Toma for a 705-type computer, will be described first. The present routine (Direct Conversion System), modified from the original by J. Moyne for a 709-type computer, will be described later in terms of its differences from the original routine.

(The original routine is everywhere described in the present tense, in part because so many of the principles carry over into the present routine, in part because continual changes in tense would make understanding difficult.

The dictionary lookup routine divides the memory of the computer into 9 main areas. These are:

1. The area for split dictionary entries;
2. The area for textwords;
3. The area for looked-up textwords to be written out;
4. The area for unsplit dictionary entires;
5. The constants;

6. The dictionary lookup routine;
7. The morphological routine;
8. IDENT; and
9. The area for patching.

1. The area for split dictionary entires.

A block of split entries is read into the area for split entries from the split dictionary tape. (There are fifty entries in each block, but the program for updating the split dictionary can organize the entries into larger or small groups depending on the capacity of this area in various types of computers.) More blocks are read in as necessary. When the current dictionary item lies beyond the point in the alphabetic sequence represented by the last item in the split entry block, the old block is removed and a new block is brought in. If, on the other hand, the first item in the split entry block lies beyond the point] alphabetic sequence represented by the current dictionary item, an old block can be returned in whole or in part.

2. The area for textwords.

A block of alphabetically sorted textwords is read into this area. There are 20 entries in a textblock. The words of one block are looked up one by one each repetition of the same word is looked up separately. When all of the wo have been looked up, the old block is removed, and a new block is brought in.

4. The area for looked-up textwords to be written out.

Each time a textword has been matched in the dictionary, both the text-word and its serial number are transferred to this area along with all of the dictionary information which follows the entry. If the word is not found, it is transferred here also with a computer-generated code (Code 8 in I:36) to indicate that it was not located in the dictionary. After all the textwords of a block has been analyzed, this area is usually full and the material in it is ready to be copied onto tape for storage. By means of a read-while-write routine, the material is copied onto a tape and, at the same time, a new block of textwords is read into the area which the lookup has emptied.

4. The area for unsplit dictionary entries.

This area is used in the same way as the area for split dictionary entries. The difference is that here the block contains only ten entries instead of fifty. This ratio between split and unsplit blocks corresponds approximately to the present ratio of split and unsplit entries in the dictionary as a whole. The same mechanisms for exchanging or recalling blocks apply here as apply in the area for split entries.

5. The constants.

The constant codes used in the program are stored in this area along with the symbolic tags which serve to identify them. The constants consist in the main of paradigmatic sets for morphological analysis and other analytic codes. The field reserved for working space is also included here as well as the codes for clearing the working area.

6. The dictionary lookup routine.

This is the set of instructions which makes up the heart of the program. It moves dictionary blocks and text blocks into the appropriate areas whenever they are required, it transfers a split entry which has been matched to the morphological routine, and it re-arranges data within the memory as required. This program will be described in a more detailed way later in this paper.

7. The morphological program.

The morphological program carries out the morphological analysis when a split form has been matched and its ending must be matched in turn. This program compares the paradigmatic set of the split form being tested with the characteristics of the particular ending which appears with it. Then the program generates the necessary morphological codes for the textword whose stem and ending have been matched. This is described in more detail in the Papers on Morphological Analysis.

8. IDENT (standard tape movement and checking).

The United States Air Force has developed a standard routine which automatically effects tape alternation, tape identification, error checking, initial reading and other similar control processes. The utilization of this program saves a considerable amount of programming and coding effort.

9. The area for patching.

A small area has been reserved for corrections and for amending the routines with additional steps.

The Lookup Routine

Standard tape movement instructions furnish the first block of textwords and the first blocks of dictionary items to the computer memory. Whenever necessary, a new block is read into the memory, or an old block is returned at least in part. There are constant checks to ensure synchronization. In the course of the following description, these tape control mechanisms will not be referred to further.

The Unsplit Dictionary

The binary numeral form of the textword is compared with the binary 9 numeral form of the unsplit dictionary entry.

There are three possible results.

1. The textword code may represent a number lower than the dictionary entry code.

   In this case the lookup moves on to the next dictionary entry.

2. The textword code may represent a number equal to the dictionary code number.

   In this case the textword has been located in the unsplit dictionary and the dictionary information which follows the entry is copied into a working area 1 where it is associated with the textword and with the serial number of that pa] ticular textword.

3. The textword code may represent a number greater than the number represented by the dictionary entry code.

In this case the textword has not been located in the unsplit dictionary and the search goes over to the split dictionary.

The Split Dictionary

The binary numeral form of the textword is now compared with the binary numeral form of the split dictionary entry. Since the dictionary entry in this case is not a complete word, while the textword is, the dictionary entry is compared only with that portion of the beginning of the textword which is equal in length to the dictionary entry. Thus a split dictionary entry STOL and a textword STOLAMI are compared by comparing the four letters of STOL with only the first four letters of STOLAMI.

When such a comparison is made, there are again three possible results.

1. The code of the first portion of the textword may represent a number lower than the number of the dictionary entry code.

In this case the lookup moves on to the next split dictionary entry.

2. The textword number may be equal to the dictionary entry number.

In this case a series of operations is initiated which identifies the textword.

a) If the word ends in the reflexive suffix (-S4, -S6), a code is generated which represents this fact, and the suffix is removed.

b) A count is made of the number of letters by which the textword now exceeds the split dictionary entry in length.

If the number of excess letters is zero, one, two, or three, the lookup transfers to the morphological routine under zero-endings, one-letter endings, two-letter endings, or three-letter endings, respectively.

If the number of excess letters is more than three, the textword is returned to the process of comparison with the split dictionary entries, because there are, at present, no morphological endings of more than three letters .

c) If the excess of the textword over the dictionary entry is identifiable with a morphological ending and if that ending is compatible with the structural role played by the split dictionary item (naturally, there must be no connecting of a morphological ending for nouns with a split entry which can only be part of a verb), the textword has been located in the split dictionary and the dictionary information and the morphological routine information are copied into a working area where they are co-ordinated. The result is then associated with the textword and with the serial number of the textword.

The textword number may be greater than the dictionary entry number.

In this case the textword is not to be found in the split dictionary. The text-word is then submitted to a number of tests, to establish whether it is an Arabic numeral, or a Roman numeral, or a formula.

a) The routine for the recognition of Arabic numerals requires that the dictionary should contain the Arabic numerals representing all integers from one to 99. All other Arabic numerals are automatically recognized and identified.

b) The routine for the recognition of Roman numerals identifies a Roman numeral from the nature of the characters which form the number, and also from a specific keypunch symbol which distinguishes these numerals.

The number and formula routine also recognizes page numbers, paren-theses, and all the notations of tables, footnotes, and so on, by the symbols which are used to identify them; the routine does not analyze them as possible candidates for number or formula transfer.

Before it is finally determined that an item is not in the dictionary, a check is made *as* to whether the word ends with -LIBO or -NIBUD6. If it does, the word is analyzed and transferred to the output. If it does not, the word is considered not to be in the dictionary.

If a textword has not been identified after all the above-mentioned steps a special routine checks as to whether the split dictionary is too far advanced. Such a situation can occur when a textword is compared with the split dictionary and is not found. In the lookup process, the split dictionary may have been moved beyond the position of the following textword in the alphabetic sequence. In such a case, the first entry of the dictionary block in the computer memory is checked.

If this first entry is anterior to the textword in the alphabetic sequence, the routine continues with the same dictionary block, but if the entry is beyond the position of the textword in the alphabetic sequence, a back-spacing takes place until the sequence positions of the split entries and textwords corresponds again in the computer memory.

When a textword remains unmatched after all the above operations, or when its grammatical function is not recognized, the routine generates the code 8 in I:36, and places the Russian word itself in the position for the English gloss the Russian word is all the information available until Gap Analysis comes into play, and so is transferred to the working area. At the same time the entry is recorded on a special tape with an indication that it was not located in the dictionary.

The Direct Conversion Lookup

The Direct Conversion Routine is essentially the same as the Serna routine. The following points may be noted.

The split dictionary is read into three areas (Al, A2 and A3) rather than into one. Each area holds 150 split items.

The unsplit dictionary is read into two areas (Bl, and B2) rather than into one. Each area holds 68 unsplit items.

The input text is read into two areas (C1 and C2) rather than into one. *M* area holds 100 sorted textwords.

There is a work area.

There is an output area, where words which have been looked up can be written out.

While the lookup process is being carried on in one set of areas, for example in A1, B1 and C1, the other areas are being cleared of blocks which

have been passed, and refilled with the blocks following those in actual use. Wo rewinding or rereading of tapes is needed except as noted in the Serna routine.

The SLC Lookup

The SLC Lookup is essentially the same as the Serna Routine, but differs from it in some respects. These differences are chiefly due to the fact that the SLC was not originally evolved specifically to handle the GAT, as both the Serna programming and the Direct Conversion Programming were. Thus, among other differences, the SLC is not limited to matching endings of three letters or less in the morphology, but can handle endings of greater length if necessary.

MORPHOLOGICAL ANALYSIS

Russian is a highly inflected language. A noun has twelve potentially different forms, an adjective twenty-four, and a verb sixty or more. In almost all cases, the beginning of the word remains the same; this part of the word is the base. The end of the word varies; this part is the ending.

Grammatical classes (nouns, adjectives, verbs, and so on) are made of different bases which take the same or similar arrays of endings.

It would require a vast dictionary to list every combination of base and ending in Russian. A much smaller dictionary is required if the bases are listed in one part of the dictionary, and the endings in another. But then it becomes necessary to have some means of associating each base with those endings which occur with it. Codes supply this means.

It is thus possible to analyze a textword by matching its first part in a dictionary of bases, and its latter part in a dictionary of endings. The grammatical function of the textword can then be discovered from a list which shows the structural value of the ending when it is associated with that particular type. of base.

This process is here called morphological analysis.

In order to facilitate morphological analysis, Russian grammatical da] are designated by a special code. This code is named the paradigmatic set, or, more briefly, the parset.

The Parset

The parset is a nine-position code.

The first position (I:36) receives the parcue, the code which designates the grammatical class (part of speech) of the entry.

Code	Designation
1	noun
2	verb
3	adjective/pronoun
4	adverb
5	preposition
6	conjunction
7	particle

(The computer-generated code 8 in this position indicates that the word was not found in the dictionary.)

The second and third positions (I:27-38) receive codes which provide further grammatical information about the entry. Since the interpretation of these codes depends on the parcue with which they occur, their description is best deferred until the discussion of each grammatical class takes place.

The fourth to ninth positions (I:39-44) receive codes which are to be interpreted in much the same way in all cases, and which have undergone a similar development. This development is best discussed here.

Originally these codings were systematically chosen. The system for nouns was as follows:

> Position 4 indicated the type of the entry, provided that the pattern of inflection followed one of the commoner patterns. Where the endings deviated from the commoner patterns of inflection, the coding in positions 5-9 indicated that fact.
>
> Positions 5-7 received codes designating cases. These codes were:
>
> | | |
> |---|---|
> | 0 | no deviation; |
> | 1 | nominative deviates; |
> | 2 | genitive deviates; |
> | 3 | dative deviates; |
> | 4 | accusative deviates; |
> | 5 | instrumental deviates; |
> | 6 | locative deviates; |
> | 7 | read next position as plural only. |
>
> The case in which the first deviation occurred was coded in position 5, the cases in which the second and third deviations occurred were coded in positions 6 and 7 respectively. If there were no further deviations and if any of these columns was still uncoded, that position received a zero. If positions 6 or 7 received a digit which represented a smaller numerical value than that represented by the digit in the preceding position, then the digit of lower value was read as referring to a case of the plural. If there might be an ambiguity, the digit 7 was coded in any one of positions 4-6 to signify that the code in the following column designated a case in the plural.
>
> In positions 5, 6, and 7, then, the cases where deviating suffixes occurred were designated. In positions 8 and 9 the deviating endings were listed. A two-character code was ascribed to each deviating suffix, and, when necessary, to each combination of deviating suffixes. The characters which occurred were the digits from 1 through 7 and #. Each single suffix had its own two-position code. Each combination of suffixes had a two-position code which was a fusion of the individual codes of the two members of the combination. In most cases the structure of the combination was transparent. In those cases where it was not, the presence of the code 3 in position 4 of the parset indicated that the code was to be read in a special way, and by this device possibilities of confusion were avoided.

The list of codes and endings follows. The cases which require special reading are marked with an asterisk.

Individual Deviations:

Code	Suffix
0#	Zero
01	A
02	E
03	I
04	U
05	4
06	EV
07	El
10	EM
11	I1
12	1
13	EE
14	UH
15	0
16	IM
17	OM
20	01
21	OH
22	AS
23	AH

Combinations of two deviations:

Code	Combined Suffixes
41	U and A
13	*EM and I
16	*EM and EV
2#	E and #
56	4 and EV
32	I and 1
31	I and I1
35	I and 4
36	EV and I
37	E and I

As previously mentioned, this system has undergone a twofold development.

First, when this operation was programmed, the programmer found that the denotations of the codes as described above were unnecessary. As long as each entry which required a different code had one, it was of no consequence how these codes were constructed. Consequently the codings in the last six positions of the parset were treated in programming as if they were completely arbitrary.

Second, it proved easier and more economical to treat a number of these deviating forms as unsplit entries, and thus place them outside the range of the morphological analysis. This was true of all the verb forms. Later, when more entries which required codings in positions 4-9 were added, the codings were added arbitrarily, the only criterion being that they should be distinctive. Thus, while many of the codings which remain still adhere to the system, there are some possibilities of the system which are no longer exploited, and there are some codes which are altogether outside the system. On the whole, it is best to regard all of these codings as entirely arbitrary.

Now that the codings in positions 4-9 may be regarded as arbitrary, the discussion of the general code system is simpler.

Nouns

Position 1: The parcue for noun is 1.

Position 2: Gender

0	any gender
1	masculine
2	feminine
3	neuter

Position 3: Animation

1	animate
2	inanimate

The following examples demonstrate the noun parset code:

DOKTOR-	111 271 001	Noun (1), masculine (1), animate (1); non-palatal stem (1); in the plural (7), the nominative (1) deviates from the regular paradigm; the suffix which deviates is -A (01).
PUT-	112 123 603	Noun (1), masculine (1), inanimate (2), palatal (1); the genitive (2), dative (3), and locative (6) singular deviate from the regular paradigm; the suffix which deviates is -I (03).

LINI-	122 336 232	Noun (1), feminine (2), inanimate palatal stem (3); the dative (3) as locative (6) singular and genitive plural (2) deviate from the regular paradigm; the suffixes which deviate are I and 1.
KREST64NIN-	111 200 000	Noun (1), masculine (1), animate non-palatal (2), no deviations from regular paradigmatic sets (this set occurs in the singular only).

Verbs

Position 1: The parcue for verbs is 2.

Position 2: Forms of the verb

1	infinitive
3	participle
4	gerund
6	finite forms

If the verb is an unsplit entry, these codes are entered by hand; if the verb is a split entry, all forms which are not participles receive the code 6, and the computer generates code 1 or code 4 to replace it as necessary, during the morphological analysis.

Position 3: Types of participles

1	past passive participle
2	past active participle
3	present passive participle
4	present active participle

m

Apart from this use in designating the type of a participle, the third I position is little used in verb coding. It receives the code A if the verb does not have a present passive participle. (This is to eliminate the possible ambiguity between the short form masculine of this participle and the first person plural of the non-past.) There is no code to distinguish present and past gerunds, although it could be added with no difficulty. Positions 4-9 originally held ending-deviation codes similar to those for nouns. They were few in number and it seemed easier to incorporate all such verb forms in the unsplit dictionary.

It should be noted that the reflexive suffix (-S4, -S6) is treated differently from other verb endings. As soon as a matching is achieved in the split

dictionary during the dictionary lookup, a search is made for the reflexive suffix. If it is found, it is removed and a code generated to indicate its erstwhile presence. Then the various morphological procedures for the verb are carried out.

Adjectives

Position 1: The parcue for adjectives is 3.

Position 2: Comparative inflection

Ø	no comparative
0	comparative possible

Position 3: Ambiguity

Ø	no ambiguity
0	ambiguity possible

The possible ambiguity is between the neuter singular nominative and accusative of the positive and the predicative form of the comparative (SINEE: 'blue'/'bluer').

Pronouns and Numerals

Pronouns and numerals are coded as special cases of adjectives.

Position 1: The parcue for pronouns, as for adjectives, is 3.

Position 2: Subclasses of Pronouns

2	possessive pronouns;
3	indefinite, relative and interrogative pronouns;
4	personal pronouns exhibiting gender;
5	definite and demonstrative pronouns;
6	personal pronouns not exhibiting gender and reflexive pronouns;
7	ordinal numbers and collective numbers.

Position 3: Gender

1	masculine only
2	feminine only
4	neuter only

/

Combinations of more than one gender are denoted by the sum of the appropriate codes, as masculine and feminine (3), all genders (7).

Adverbs

The parcue for adverbs is 4. Positions 2-9 are zeroed in.

Prepositions

The parcue for prepositions is 5. Positions 2-9 are zeroed in.

Conjunctions

The parcue for conjunctions is 6. Positions 2-9 are zeroed in.

Particles

The parcue for particles is 7. Positions 2-9 are zeroed in.

Matching of Endings

All of the endings needed for morphological analysis are divided into *it* groups according to their length in letters; this division speeds the ending-matching routine.

There is one ending in the zero group. There are thirteen in the one-letter group. There are fifty-three in the two-letter group and thirty-three 9 in the three-letter group. No ending has more than three letters.

When the textword has been matched with an entry in the unsplit dictionary the excess of letters of the textword over the entry is counted and is then matched in the dictionary of endings of that length. When the ending is matched, the parset of the entry is compared with the list of parsets which are compatible with that ending. When a match is found, information as to the codes which 1 are to be generated is retrieved.

Suppose a noun base with parset 1221ØØØØØ is found with the ending -I. Since -I consists of one letter, a search is made in the list of one-letter ending and an ending -I is located. Then a search is made to discover if the ending is compatible with the parset 1221ØØØØØ. The partial table below shows a possible arrangement of the information.

ENDING -I:

Noun Parsets	Cases of Singular	Cases of Plural
1111ØØØØ	.....	1 .....
111172ØØ6	.....	1 .....
11211ØØØ	.....	1..4..
112172ØØ6	.....	1..4..
1221ØØØØ	.2....	1..4. .
122172ØØ7	.2....	1..4..
132271ØØ3	.....	1..4..
132367231	6	.....

From this table it appears that the parset and the ending are compatible, and that the combination is genitive singular (2), or nominative (1) or accusative (4) plural.

Examples of Morphological Analysis

Nouns:

A textword which is a noun is compared with the split entries and matched. It is then compared with the endings and matched. The computer thereupon generates codes for case and number. (Gender is permanently coded for most nouns; surnames, such as IVANOV and ZELENSKI1 are treated as nominals but are coded in the dictionary as being of no gender, because the gender cannot be determined from the stem itself; in such a case, the computer generates a code for gender also.)

Examples: KNIGO1 is analyzed by matching first with the base KNIG-, and then with the ending -O1. This yields the information: instrumental singular. The computer stores the appropriate code in the appropriate location (5 in III:70).

STOL is analyzed by matching first with the base STOL- and then with the zero ending. This yields the information: nominative and accusative singular. The computer stores 1 in III:66 and 4 in III:69.

ZELENSKI1 is analyzed by matching the base ZELENSK- with the ending -I1. This yields the information masculine nominative singular (1 in III:79 and 1 in III:66).

Adjectives:

A textword which is an adjective is compared with the split entries and matched. It is then compared with the endings and matched. The computer

generates codes for gender, case, number, degree of comparison and form

Examples: XOROWIM is analyzed by matching first with the base XOROW-, and then with the ending -IM. This yields the information: instrumental singular (5 in III:70), dative plural (3 in III:74), all gender (7 in III:79), positive degree (no codes), long form (1 in III:80).

NOVI is analyzed by matching first with the base NOV- and then with the ending -I. This yields the information: nominative plural (I in III:72), masculine (I in III:79), positive degree (no codes), short form (2 in III:80).

Verbs:

A textword which is a verb is compared with the split entries and matched It is then compared with the endings and matched. The computer generates codes for tense, number, person, voice and gender as applicable.

Examples: CITAHT is analyzed by matching first with the base CITA- and then with the ending -HT. This yields the information: plural (2 in III:78), all genders (7 in III:79), active voice (2 in IV:37), non-past tense (2 in IV: 39) and third person (3 in IV: 38).

SOBIRALS4 is analyzed by matching first with the base SOBIRA-, removing the reflexive ending -S4, and then by matching with the ending -L. This yields the information: singular (1 in III:78), masculine (1 in III:79), reflexive (3 in I:68), past tense (1 in IV:39) and all persons (6 in IV: 38).

Sometimes two stems have the same form. The verb RASTVOR-IT6 has the same stem as the noun RASTVOR-Ø; the noun DN-O has the same stem as the noun DN-4 (genitive).

The computer will normally search only as far as the first possible listing encountered unless there is some indication that another possibility exists. The code which gives this indication is N in I:34.

Example: The textword RASTVORIM is matched against the stem RASTVOR- (noun; parset 1122ØØØØØ). The suffix- matching routine finds no listing for a suffix -IM with a noun stem of the type 1122ØØØØØ. Ordinarily, the computer would generate a code 8 in parcue to indicate that the word is not in the dictionary. But first, a search is made for the code N in I:34, and when it is found, the search is continued to the next stem RASTVOR (verb; parset 26). The suffix matching operation is successful; the computer generates codes for first person (1 in IV: 38), plural (2 in II:78), non-past tense (2 in IV: 39), active voice (2 in IV: 37).

MORPHOLOGICAL ANALYSIS

Alternant Bases (Verbs)

When forms are split for entry in the split dictionary, only one cut is made in each form. The location of the cut is justified in terms of machine translation only. A verbal base in machine translation corresponds to either a root or a stem in conventional grammar, depending on the class of verb involved. The reason for this variation is that the stem-forming vowel is assigned sometimes to the base and sometimes to the ending. The variability is justified in terms of computer operations; the location of the cut is determined on the basis of economy and simplicity; the chief purpose is to reduce the number of dictionary entries to a minimum, without removing any of the combinatory possibilities of bases with sets of endings.

Verb bases which are subject to morphophonemic alternations were originally entered in the dictionary as split entries. Since there are usually two alternate bases in all such verb paradigms, two split entries were necessary. For example, the verb PISAT6 (to write) was entered twice. It was entered under PIS-, which combines with the endings -AT6, -AL, etc., and it was entered under PIW- which combines with the endings -U, -EW6, etc.

The number of verbs which exhibit morphemic alternation is relatively high. It therefore seemed advisable to develop a procedure which would permit of handling alternate verb bases as one single entry instead of as two or more entries. Such a procedure would significantly reduce the number of dictionary entries. It also seemed advisable to classify the verb bases on the basis of their distribution, since each base alternant occurs with a limited and specific array of endings; such as classification would reduce the number of possible ambiguities which might arise from mismatching.

The following system has been developed but has not yet been programmed.

Patterns of Alternation

A study of the differences in the alternant forms has been made, and, on the basis of the letters involved, thirty-eight different patterns of paired alter- have been established. These thirty-eight sets of paired alternants fall into three major categories, depending on the number of letters at the end of the base which are different in the alternant forms. In one alternant of every pair, the alternation pattern affects only the one last letter of the base, and this alternant is the form used as a dictionary entry. Thus, every alternant is assigned the number 1 as the first part of its designation. In the other alternant of any pair, the alternation pattern affects the last one, two, or three letters of the base. Thus, the alternant is assigned the number 1, the number 2, or the number 3 as the second part of its designation. Any pair of alternants is designated, under this system, as a 1-1 alternation if the difference is only in the last letter of

each alternant, as a 1-2 alternation if the difference is in the last letter of one alternant as against the last two letters of the other alternant, and as a 1-3 alternation if the difference is in the last letter of one alternant as against the last three letters of the other alternant.

The thirty-eight patterns of alternation can be sub-divided as follows:

a) 1 - 1 alternations (24 patterns),

b) 1 - 2 alternations (12 patterns),

c) 1 - 3 alternations (2 patterns).

Under this system, morphemic alternations are described only when the final of the base is involved in the alternation. Changes in the interior of a base still require that two or more bases be listed.

Alternants which are the result of suppletion are entered in the dictionary as unsplit forms for simplicity of programming.

Alternation Code

The coding which signals the different patterns of alternation utilizes four positions.

A digit in the first position indicates the part of speech; since the discussion here covers verb bases only, only the code 2 occurs in this position in this paper.

If a verb has only one base, the second, third and fourth positions are zeroed in. If a verb has alternant bases, the codings in the second, third, and fourth positions spell out the letter which must replace the last letter of the dictionary entry in order to produce the alternant. It is particularly to be noticed that the stem alternant selected as the dictionary entry is always such that only the one last letter needs to be dropped before the one, two, or three letters which produce the other alternant are added.

An example of a verb with 1-1 alternation is the verb PISAT6: 'write'. It is entered in the dictionary as: PIS- 2WØØ. The code W shows that the final S of the entered stem (alternant 1) is replaced by W in the other alternant.

If the textword is PIWET, it will not be matched immediately in the look-up with the entry PIS-, but, since PIS- is coded to allow also of the alternation PIW-, a matching is made on this basis and the lookup goes on to the routine for the suffix -ET and establishes whether the suffix -ET does occur with the alternant PIW-. Since it does, the matching is complete.

An example of a verb with 1-2 alternation is the verb RISOVAT6: 'draw'. It is listed in the dictionary as RISU- 2ØVØ. The one-position final U of alternant 1 alternates with the two-position final OV of alternant 2. Into the same category fall other 1-2 alternation types such as JEVAT6 (JU- 2EVØ) and PLEVAT6 (PLH- 2EVØ).

An example of a verb with 1-3 alternation is RAZBIT6: 'break up'. It is listed in the dictionary as RAZB- 2OB6. The one-position final B of alternant 1 alternates with the three-position final OB6 of alternant 2.

The introduction of a zero functioning as the final of the dictionary entry (alternant 1) makes it possible to code many types which could not otherwise be treated directly by the process described above.

Verbs of the type GASNUT6 are listed as GASØ- 2NØØ. The extension of the base by adding zero will result in the following operations for endings.

GASØ- Ø; LA; LO; LI.

GASN- U; EW6; ET; EM; ETE; UT.

Verbs such as JEC6 must be entered in the dictionary in many forms: JEC6 and JEG are entered in the unsplit dictionary, while the stem JG- is entered in the split dictionary and has an alternation code 2JØØ.

JG- U; UT; LA; LO; LI.

JJ- EW6; ET; EM; ETE.

Two separate entries are still required for verbs such as POSLAT6: 'send' and MOLOT6: 'grind', since the alternation cannot be achieved by removing one letter from the end of either base in either verb. (POSL-/POWL-) (MOL-/ MEL-).

All anomalous verb-forms, such as those of EST6: 'eat', or of ITTI: 'go', are listed in the unsplit dictionary.

Here are assorted examples of verb listings.

Infinitive	Dictionary Entry		Alternants	
			First	Second
XOTET6	XOT-	2CØØ	XOT-	XOC-
VLEC6	VLEK-	2CØØ	VLEK-	VLEC-

Infinitive	Dictionary Entry		Alternants First	Second
NOSIT6	NOS-	2WØØ	NOS-	NOW-
KLAST6	KLA-	2DØØ	KLA-	KLA-
MYT6	MY-	2OØØ	MY-	MO-
PIT6	PI-	26ØØ	PI-	P6-
JAT6	JA-	2NØØ	JA-	JN-
JAT6	JA-	2MØØ	JA-	JM-
TERET6	TR-	2ERØ	TR-	TER-
BRAT6	BR-	2ERØ	BR-	BER-
GNAT6	GN-	2ONØ	GN-	GON-
STLAT6	STL-	2ELØ	STL-	STEL-
ZVAT6	ZV-	2OVØ	ZV-	ZOV-
ISKAT6	I5-	2SKØ	I5-	ISK-
RAZBIT6	RAZB-	2OB6	RAZB-	RAZO-

Array Codes

Anyone familiar with these Russian verbs will quickly realize that the array of the endings which occur with the first and the second alternants is no means the same in all verbs. A few examples will suffice to illustrate this fact.

Verb	Alternant	Array of Endings
BRAT6	BR-	-AT6, -AL, -ALA, -ALO, -AL
	BER	-U, -EW6, -ET, -EM, -ETE, - - UT
TERET6	TR-	-U, -EW6, -ET, -EM, - ETE, -UT
	TER-	- ET6, -LA, -LO, LI
NOSIT6	NOS-	-IT6, -IW6, -IT, -IM, -ITE, -4T, -IL, -ILA, -ILO, -ILI
	NOW-	-U

Accordingly, provisions must be made to avoid mismatchings of endings to be with which they do not occur.

Each array of endings which occurs with any base is listed. Arrays which are the same in all respects are considered to be the same array.. Each array is assigned a code which identifies it.

Each verb alternant receives the appropriate array code.

The verb endings are listed individually in the dictionary of endings. They are associated in three groups, according to their length in letters. The total number of one-letter endings for verb bases is 9; the total number of two-letter endings is 20; the total number of three-letter endings is 26.

Under each ending in the dictionary of endings are listed the codes of all arrays which include that ending. With each array code is given a list of the morphological values associated with that ending in that array.

In this way, when a verb alternant has been matched, the array code associated with it is noted. When the ending is matched, a search is made under that ending for the array code of the alternant. If it is found, a listing of the morphological values of that ending in that array is found, and these morphological values are entered in the appropriate positions of the working material as computer-generated codes.

Thus, NAPIWITE is not matched in the unsplit dictionary, and is not matched directly in the split dictionary. But the split dictionary entry NAPIS- (:'write') shows a code which permits of NAPIW- as an alternant stem, and the matching is made. The array code with NAPIW- is, perhaps, AQ. The excess of the textword over the split dictionary entry is three letters. A search is made among the three-letter endings for an ending -ITE, and such an ending is found. A search is made for an array code AQ under -ITE, and such a code is found. The information found in connection with this array code is that the morphological value of the ending is 'second person plural imperative'. .

Similarly, NOSITE is not matched in the unsplit dictionary, but it is immediately matched in the split dictionary with NOS- (: 'carry'). The array code is, perhaps, AD. Under the three-letter ending -ITE is found the array code AD with the information that the morphological value is 'second person plural either non-past or imperative'.

Again, SVITE is not matched in the unsplit dictionary, but is immediately matched in the split dictionary with SV- (: 'wind up'). The array code is, perhaps, AK. Under the three-letter ending -ITE there is no array code AK. Therefore SVITE is not connected with the verb base SV- : 'wind up'. (It is the dative or locative singular of SVITA: 'suite'. In the original system of morphology, the possibility exists that SVITE would be recognized as some form of the verb SVIT6.)

This system of handling the verb morphology will be more compact than the original system. It will also be more precise, since the system of array codes reduces the possibilities of mismatching.

MORPHOLOGICAL ANALYSIS

Complementary Distribution in Endings (Nouns and Adjectives)

A second approach to the morphological analysis of Russian nouns and adjectives has also been proposed.

The first approach (the original morphological analysis) classifies declen -sional forms purely on the basis of their spelling in the GAT transliteration.

The second approach classifies forms not only on the basis of their spelling in the transcription, but also on the basis of sets of inflectional endings.

The sets of endings are so chosen that no member of any set occurs with the same base as any other member; the endings of each set are in complemen- tary distribution; each set can be regarded as a unit.

In this way, the number of separate endings needed in the dictionary can be substantially reduced. At the same time, the number of classes of nominal and adjectival bases can also be substantially reduced. The traditional gram-matical distinction between so-called "hard" and "soft" declensions is eliminated For example, in the first approach, STOL, FLAG, KORABL6, MUZEl NOJ and PALEQ are separate declensional classes which combine with different arrays of single endings. In the second approach, they are united in one class which combines with one array of sets of endings. This array is the following.

nom. and acc. sg.	-Ø/-6/-1
nom. and acc. pl.	-Y/-I gen. sg. -A/-4|
gen. pl.	- ØV/-EV/-E1
dat. sg.	-U/-H
dat. pl.	-AM/-4M
instr. sg.	-OM/-EM
instr. pl.	-AMI/-4MI
loc. sg.	-E
loc. pl.	-AX/-4X

Each set of endings has a general value.

The general value of the set of endings is the sum of the structural value of that set of endings, plus the sum of its distributional values.

For example, the general value of the set of endings -A/-4 is the sum of the structural values nominative, genitive and accusative singular, and nominative and accusative plural, plus the sum of the distributional values of its combining with nouns of the types VODA, ZEML4, etc. , STOL, KORABL6, MUKRAL6, PROFESSOR, etc., SELO, POLE, etc.

The general value is a constant. As such it can be recorded in the dictionary of endings. However, the use of the general value may result in the generation of wrong information, since no ending has its general value with all bases, but has a particular value with each base.

Some mechanism must be established in the dictionary which will induce the computer to select the particular value of the set of endings appropriate to the base with which it occurs.

Particular values are either unambiguous in number and case or are ambiguous.

It follows that three classes of ambiguous values can exist; in Russian, only two classes actually occur.

Class A contains those patterns where both number and case are ambiguous.

Class B contains those patterns where case is ambiguous but number is unambiguous.

Class C would contain those patterns where number is ambiguous but case is unambiguous; there are no such patterns in Russian.

Subclassification of the ambiguous pattern classes is based on the number of possible ambiguities in that subclass. Each subclass is here designated by a numeral which represents the number of possible ambiguities.

In Class A, the following subclasses are found:

Subclass 2: loc. sg. ; nom. pl.	PROLETARI-I
gen. sg. ; nom. pl.	JEN-Y
instr. sg. ; gen. pl.	DOL-E1
instr. sg. ; dat. pl.	IVANOV-YM
Subclass 3: instr. and loc. sg. ; dat. pl.	VS-EM
gen. and acc. sg. ; nom. pl.	DOKTOR-A
gen. sg. ; nom. and acc. pl.	OSTROV-A
loc. sg. ; nom. and acc. pl.	SANATORI-I
instr. sg. ; gen. and acc. pl.	QAPL-E1
nom. sg. ; gen. and acc. pl.	SOLDAT-0
nom. and acc. sg. ; gen. pl.	GLAZ-0
Subclass 4: nom. and acc. sg. ; nom. and acc. pl.	VS-E
gen. , dat. and loc. sg. ; nom. pl.	MYW-I
Subclass 5: gen. , dat. and loc. sg. ; nom. and acc. pl.	REAKQI-I
Subclass 12: all cases both singular and plural.	TAKSI

In Class B, the following subclasses are found:

Subclass 2:	gen. and dat. sg.	SAXAR-U
	nom. and acc. sg.	STOL- Ø
	gen. and acc. sg.	SLON-A
	dat. and loc. sg.	LES-U
	nom. and acc. pl.	STUL6-4
	gen. and acc. pl.	SLON-OV
	instr. and loc. sg.	C-EM
	dat. and acc. sg.	IVANOV-U
Subclass 3:	gen. , dat. and loc. sg.	BEDNOST-
	nom. , acc. and loc. sg.	POL/-E
	nom. , gen. and acc. sg.	IVANOV-A
	gen. , acc. and loc. pl.	IX
Subclass 4:	gen.,dat., instr. and loc. sg.	ST-A
Subclass 6:	nom., gen., dat., acc., instr. and loc. sg.	KOFE

The particular structural value of a set of endings, whether ambiguous or unambiguous, is derivable by a process of selection from the overall value The particular value is a variable, since the selection depends on the stem which any member of the set is combined.

The possible selection patterns of structural values are never many.

In theory, there are 3125 (5^5) potential selection patterns for the structural values of the set of endings -A/-4; in practice, only nine patterns occur

a)	gen. and acc. sg.	SLON-A
b)	gen. and acc. sg. ; and nom. pl.	DOKTOR-A
c)	gen. sg. , nom. and acc. pl.	SLOV-A
d)	nom. and acc. sg.	VREM-4
e)	nom. and acc. pl.	VREMEN-A
f)	gen. sg.	STOL-A
g)	nom. sg.	JEN-A
h)	nom. pl.	BRAT6-4
i)	gen. sg. , and nom. pl.	LIQ-A

Some mechanism must be established in the dictionary which will induce the computer to select the appropriate particular value in each case. Array codes achieve the desired effect.

Each different array of sets of endings which occurs with a base in declension receives an identification code, and this code is entered in the dictionary against each nominal or adjectival base.

Under each ending in the dictionary of endings is a reference to the set of endings to which that ending belongs. Under each set are listed the codes of all arrays which contain that set of endings. Against each array code is listed the particular structural value of that set of endings in that array.

This information is copied into appropriate positions of the working material as computer generated codes.

For example, STOLA is not matched in the unsplit dictionary, but it is matched in the split dictionary. The split dictionary entry STOL- (: 'table') carries an array code, say NP. The excess of the textword over the dictionary entry is one letter. A search in the dictionary of endings shows that -A is a possible one-letter ending, and that it belongs to the set -A/-4. Under the set -A/-4 are listed the array codes of all arrays which contain the endings -A or -4. A search is made for the array code NP. It is found and, with it, the information 'genitive singular'. The computer generates codes to indicate that STOLA is genitive singular.

It will be observed that the second approach to noun and adjective morphology is similar in a number of points to the approach to verb alternants, and includes the same safeguards against mismatching.

Special care must be taken, however, in the case of homographic stems. Thus DNO (: 'bottom') has the genitive DNA, and DEN6 (: 'day') has the genitive DN4. To a system of morphological analysis which considers the endings -A and -4 as being in complementary distribution, DNA and DN4 are the same form. Therefore the array code for DN- from DNO must indicate clearly that DN4 is not acceptable as a genitive form, and the array code for DN- from DEN6 must similarly indicate that DNA is not acceptable as a genitive form.

A Comparison of the Two Approaches

The second approach to noun and adjective morphology is uni-directional and can be used only if Russian is the source language. There is the advantage that the number of classes of bases is substantially smaller than it is in the first approach, and that there are more mechanisms for ensuring correct matching of base and ending.

The first approach is ambi-directional. It can be used if Russian is either the source or the target language. There is the advantage that the number of stem subclasses is substantially higher than in the second approach, since the various forms of each set of endings must be treated as separate units.

The user will of course decide which of these approaches is more advantageous for his purpose.

MORPHOLOGICAL ANALYSIS

Base Analysis

The term base analysis may be applied to the process of analyzing bases to establish their components. The morphological analysis of the GAT analysis of words into base and ending; base analysis then analyzes the bases into further components. Thus PRIVODA is analyzed into the base PRIVOD- and the ending -A by the morphological analysis. It can be further analyzed into a prefix and a root -VOD- if this seems desirable, and this procedure would fall into the realm of base analysis.

In general in the GAT, base analysis is not widely used. A modified base is used in the Gap Analysis which is described in a later paper of this report. The chief area in which base analysis has been accorded importance in the GAT is in dealing with chemical terminology.

Chemical Terminology

The terminology of organic chemistry provides a large number of text items which are names of chemical compounds. These items have particular characteristics. First, their number is extremely great. Second, each compound is named according to established rules and with the aid of a restricted number of components, which have been referred to as "chemical morphemes". Third, these components belong to the international scientific vocabulary, and show a great deal of similarity, both in form and in manner combining among themselves, whether in Russian or in English.

If the name of each organic compound were entered in the dictionary as a separate entry, the size of the dictionary would increase rapidly, and would become unwieldy.

But since these names are composed of a restricted number of components (the chemical morphemes), it is possible to contrive a limited split chemical dictionary where the entries are base components rather than entire bases.

An attempt to contrive such a dictionary was made. The operation was designed, but the computer routine was never programmed, and so the operation has not been tested. The operation is as follows.

Each Russian chemical term is held to consist of three possible levels

a) the post-final, which is any component, most frequently a number, that occurs after the morphological ending,
b) the morphological ending, and
c) the base.

The operation checks all words which have not been matched in the main dictionary.

The item is checked first for post-finals. The post-finals are conventionally single digits. (They are always keypunched as a comma or a dash followed by the digit, however, because, in the GAT transcription, some digits (1, 3, 4, 5, 6, 7) represent Russian letters; the prefixed dash or comma is necessary to distinguish a letter which is symbolized by a digit and which is a part of the stem from a digit which is to be read as a post-final.) The post-final is compared with a list of all possible post-finals. When a post-final is matched, its value is copied into the area for the English meaning, and a search is made for another post-final. When no further post-final is found, the operation proceeds to the analysis of base and ending.

The segment containing the base and ending is compared, from the right, with the entries in a special dictionary of base components.

These base components are entered in their dictionary so that the morphemes which are longest in terms of the number of letters are listed first and the shorter morphemes later; morphemes of the same length are listed according to the alphabetic order of their final letter. If a final stretch of the item is matched with one of the base components, it is assumed that the inflectional ending is zero; the parcue of the base component is checked to see if it is compatible with a zero ending and, if so, the morphological value of the combination is established.

If no match is made, the operation shifts one letter to the left, and a comparison is made of all but the last letter of the item with the base component dictionary. If a match is made, it is assumed that the inflectional ending is one letter long (the last letter that was not matched); the parcue of the base component is checked to see if it is compatible with that inflectional ending, and if so, the morphological value of the combination is established.

If no match is made, the operation shifts one more letter to the left, checking the possibility that there is a two-letter inflectional ending. If there is still no matching, the operation shifts one more letter to the left, checking the possibility that there is a three-letter inflectional ending. No inflectional ending is longer than three letters, and, if no match has been made by this point, no match is possible.

When a match has been made, and the morphological value of the inflectionnal ending has been established, the still unmatched portion of the item is compared with the base component dictionary, again from the right until a match is made. This process is repeated until no residue remains.

As each match is made, the gloss of the segment matched is copied in strict order from the right into the area for the English translation. The sum of the glosses of the segments serves as the gloss of the whole.

The entire chemical term and its gloss can then be stored temporarily and, if there is more than one occurrence of the term in the test, the information which has already been generated can be simply copied for each occurrence there is thus no need to generate the same information again for that text.

Consider 4-GALOGEN-2-XLORBUTENA-3 as an example both of the general effectiveness and of a particular deficiency of this operation. The final -3 will be matched first. Its gloss, -3, will be stored at the right of the area for the English meaning. There should be no matching of XLORBUTE as it stands with any of the base segments, and the operation should shift one position to the left and proceed to compare XLORBUTEN with the base segment. There will be a match with EN and the gloss 'an(e)' will be read into the area for the English meaning immediately to the left of '-3'. (The (e) on this gloss indicates that the e is to be used if the segment is the first one matched, as this is.) The -A will be read as an inflectional ending one letter in length and the morphological analysis will show it as genitive singular. Subsequent matching will produce the glosses BUT : 'but', XLOR : 'chloro', -2- : '-2-GALOGEN ; 'halogen' and 4- : '4-'. The translation should appear as '4-halogen-2-chlorobutane-2' and the form should be noted as being genitive -. singular.

Unfortunately, A is listed as a chemical morpheme with the gloss 'a'. and so the desired translation is not generated in this particular case, but '4-halogen-2-chlorobutana-2' is generated instead. This type of deficiency while not by any means ubiquitous, must nonetheless be eliminated before I operation's output can be considered acceptable. Some method of marking those base components which are ambiguous with morphological endings is needed.

A subsequent attempt at a solution was made by Dr. Lawrence Summer a chemist from the University of North Dakota, who spent a year at George studying the machine translation system. The general principles of the word analysis remain the same in this new system, but there are certain marked differences. The comparison begins from the left instead of from the right. Certain chemical morphemes were marked as having alternant forms in com-position; thus, XLOR(O) represents either XLOR or XLORO. Other chemical morphemes were marked as having a restricted distribution; thus ID: 'id(e) is marked as never occurring initially. A special lookup procedure is suggested for the unmatched remainder of the form once this is reduced to six letters or less. The details of Dr. Summer's operation have been published The Journal of Chemical Documentation, 2, 83 (1962).

Further Research

As interest in other than chemical texts increased, the possibilities of reducing dictionary size by base analysis of certain areas of the general vocabulary began to be explored. A report on the Transfer of Russian Words of Foreign Origin has been prepared; (see the list of terminal reports).

The correspondences between English words ending in certain Latinate affixes such as -ion, -ation, -ate and Russian words ending in Latinate suf-fixes such as -I4, -AQ14, -IZIROVAT6 were studied to determine whether satisfactory transfer could be achieved by transferring the suffixes only, without the necessity of transferring the stems (MOTIV-IROVAT6: 'motiv-ate').. While the results are very interesting, the work needs more research, especially in the field of semantics, before it becomes decisive enough to show whether any significant dictionary reduction can be achieved by this method.

Base Analysis in the French-to-English Translation

In the French-to-English translation, a similar procedure for base analysis can be used if desired.

If a French word is not found in the dictionary, its termination can be compared with the entries in a special dictionary of endings. If a match is made, a certain amount of grammatical information can be generated, and, in cases, a translation of the word can be synthesized on the assumption that any word which is rare enough not to be in the general dictionary is probably international enough to have a direct counterpart in English.

Assume that the word 'polymeriserait' is not found in the dictionary. A comparison with the entires in the dictionary of terminations gives matchings for the portion '-iserait'. These matchings provide the information that the form is third person singular conditional of a verb which probably corresponds to an English verb having the suffix '-ize'. It is assumed that the form may be transferred as 'would polymerize', and this assumption is, in fact, well founded.

On the whole base analysis seems to have a value only in the case of highly technical international vocabulary where very little variation in semantic range is to be expected.

THE FORM OF THE LINGUISTIC STATEMENTS

At points in the succeeding text, linguistic statements are presented in form which has been referred to as a "verbalized flow-chart", and which evolved originally for the English-to-Turkish Translation Project. This form has the advantage that it can be written out on a typewriter, and need not be drafted, as flow-charts generally are.

A statement may be in one or more sections; each section is identified by a capital letter.

Each section is composed of a list of tests and commands. The items of the list are identified by serial numbers.

Each test is to be made as indicated; there is the possibility of either positive or a negative result. Each command is simply to be obeyed as indicated.

To the right of the list of tests and commands are three columns, headed Y, N, and A. The Y stands for 'yes' and indicates the course to be followed if the result of a test is positive. The N stands for 'no' and indicates the course to be followed if the result of a test is negative. The A stands for 'afterwards' and indicates the course to be followed when a command has been obeyed. A number in a column after a test or command indicates a movement to the test or command in the same section which is identified by that number. A letter in a column after a test or command indicates a movement to the beginning of the section which is identified by that letter.

Certain conventional signs are used in the columns. Those occurring in this report are these. A zero (Ø) in either the Y or the N column indicates that that specific result of the test has not been explored, usually because nothing in any of the texts has as yet required that it be explored. The number 99 in the column indicates that the operation is ended. An asterisk in a column indicates that the result is logically impossible, and that some error has therefore been made; a tightly-written statement will ordinarily never require the use of the asterisk, but it is useful while the statement is still in preparation.

Certain symbols are used in the tests and commands. The symbol i indicates the linguistic item under consideration at that specific point in the operation. An operation which is to deal with prepositions will usually begin with the command: Let the preposition be i.

The symbol + indicates movement or position towards the right in the Thus i+1 represents the first item to the right of the item under consideration Conversely, the symbol - indicates movement or position to the left in the text Thus i-3 represents the third item to the left of the item under consideration.

The symbol n has a numerical value which is indefinite when n is first used (usually for searching for a specific form), but which may become definite (if the specific form is found). The symbol n retains its definite value only within the section in which it occurs; the symbol +n is not necessarily of the same magnitude as the symbol-n in the same section. Thus, if it is desired to find a noun to the right of the preposition i, the test is formulated as: Is there a noun i+n? The first noun to the right will produce a 'yes' answer; if this noun is in position i+4, then +n will have the value of +4 for the remainder of the section dealing with that noun.

The symbol a has a numerical value which is fixed by a command in the operation, and which may be changed to another definite value at any point. The change of value is usually an increase or decrease of one, since the chief use of this symbol is to ensure the application of the same test to items in series under restrictive conditions. Thus, if it is desired to locate a noun object to the right of the preposition i, but only under the condition that all intervening items are to be adjectives or Class VI adverbs, part of the operation will have some such form as this.

C	Y	N	A
1. Let a be 1.	-	-	2
2. Is i+a a noun?	6	3	
3. Is i+a an adjective?	5	4	
4. Is i+a an adverb of Class VI?	5	D	-
5. Increase a by 1.	-	-	2
6. Mark i+a as the object of the preposition.			99

SENTENCE SEPARATION

The sentence separator operation was originally intended to be a part of the syntactic operation. It became clear, however, that the operations which precede the syntactic operation would become more effective if sentence separation had already been accomplished. For this reason, the sentence separation was moved forward in the series until it came to stand immediately after the morphological analysis.

In the Direct Conversion programming, the sentence separator operation has been divided into a number of segments; these segments were attached to the sentence buildup routine (a mechanical routine which immediately follow the morphological routine), to the interpolation routine, and to various routines between these two. As a result of certain changes, much of the linguistic value orientation of the sentence separator operation has been altered, and the values of sentence separator routines are now largely mechanical programming devices

In the SLC programming, the sentence separator routine precedes all other routines and retains its original character, which is described below. The SLC program does not necessarily analyze machine sentences individually however, but may cross the machine sentence boundaries under specific circumstances.

Separation

The sentence separator operation segments each text sentence, where necessary, into machine sentences (clauses); each machine sentence should then contain one basic syntactic structure which can be treated as a self-contained structural unit. (A basic syntactic structure in Russian consists a subject (H) and a predicate (P); the details are given in the paper on Syntax. In most cases, the machine sentences produced by the sentence separator operation do contain one basic syntactic structure, but some may contain two basic syntactic structures so fused that a special operation in the syntactic analysis is required to distinguish them.

Sentence separation depends upon two types of cues. One type of cue is punctuation (a). The other is the identity of specific words and their environ-ments; these words are usually conjunctions (b).

a) For the purpose of the separator operation there are two types of punctuation: single limits (or terminal punctuation) and paired limits (or parenthetical punctuation). Single limits (semi-colon, colon, single dash) are points in the text which indicate that there is one syntactic unit to the left and another to the right. Paired limits (quotation marks, parentheses, paired dashes, brackets) are pairs of points in the text which indicate that there is one syntactic structure between the two points and another which potentially extend

both to the left of the point on the left and to the right of the point on the right. Some items of punctuation, such as the comma, function both as a single limit and as a paired limit.

b) The specific words which serve as sentence separator cues are best defined by list. The list is given in the following discussion of codes.

Separator Codes

Separator codes mark those items which serve as sentence separator cues. These codes are one-position alphabetic codes, and are entered in I:71. The codes are conditioned codes; when they are sensed, tests must be made to ascertain whether certain other conditions are present; if these conditions are present, the code is valid for the separator operation.

Code	Items Receiving Code	Conditions under which code is valid
A	Colons, semi-colons	Under all conditions.
A	CTO	If i-1 is a comma.
B	GDE, ZATEM, KOGDA, ODNAKO, P03TOMU, PRICEM, TO, XOT4, CEM, any form of KOTOR-	If i-1 is a comma.
C	E., JE, TO, TOL6KO, CEGO, CTO, any form of KOTOR-	If i-2 is a comma.
D	CEGO, and any form of KOTOR-	If i-3 is a comma.
E	any form of KOTOR-	If i-4 is a comma.
F	BOLEE, UJE, any reflexive verb	If i-n is KAK and if i-1 is a comma.

K

is a code assigned to CTO to subsume codes A, B, C.

is a code assigned to KOTOR- to subsume codes B, C, D, E.

is a code assigned to TO to subsume codes B, C. is a code

assigned to CEGO to subsume codes C, D.

The Operation

Each item of the Russian text is tested first to determine whether it is a mark of terminal punctuation or an item in an excluded stretch.

If it is neither, a test is made to determine whether it is one of a number of conjunctions. Certain conjunctions, such as ESLI, KAK, NO, have tests applied which establish whether a syntactic structure is introduced by that conjunction and whether it is terminated by a mark of punctuation to the right of the text. A mark of sentence separation is placed at the mark of punctuation if it is recognized as the boundary of a basic syntactic structure.

If the item under consideration is a comma, tests for immediately following conjunctions or for certain particles or relative pronouns to the right establish whether a mark of sentence separation is to be placed at that comma.

The operations may be summarized as follows:

			Y	N
A.	1.	Is i a mark of terminal punctuation?	99	2
	2.	Is i in an exclusion area (code X in I:48)?	25	3
	3.	Is i the word ESLI?	4	5 1
	4.	Is there an infinitive i+n?	10	13
	5.	Is i the word KAK?	6	13
	6.	Is i+1 a short-form participle ?	10	7
	7.	Is i+1 a verb?	10	8
	8.	Is i+2 a short-form participle?	10	9
	9.	Is i+2 a verb?	10	11 1
	10.	Is there a comma i+n?	22	13
	11.	Is there a comma i+n?	12	B
	12.	Does i+n+1 have the code N or the code O?	22	B
	13.	Does i+n have the code A?	22	14
	14.	. Does i+n have the code K?	23	15
	15.	Is i+n an opening parenthesis?	16	17
	16.	Is i+n+m a closing parenthesis?	21	Ø
	17.	Is i+n the word NO?	18	B
	18.	Is i+n+2 a comma?	19	B
	19.	Does i+n+3 have the code B (or the code N)?	20	B
	20.	Is i+n+m the word ODNAKO?	24	B
	21.	Place a sentence separation boundary at i+n+m.		
	22.	Place a sentence separation boundary at i+n.		
	23.	Place a sentence separation boundary at i+n-1.		
	24.	Place a sentence boundary at i+n+2.		
	25.	Let i+1 be i.		-

B.

		Y	N	A
1.	Is i+n a comma?	2	18	-
2.	Let this comma be i.			3
3.	Is i+1 the word ESLI?	4	6	-
4.	Is there an infinitive or a finite verb i+n?	5	19	-
5.	Is i+n+m a comma?	16	19	
6.	Is i+1 the word KAK?	7	13	-
7.	Is i+2 a short form participle?	11	8	
8.	Is i+2 a finite verb?	11	9	-
9.	Is i+3 a short form participle?	11	10	
10.	Is i+3 a finite verb?	11	9	
11.	Is there a comma i+n?	17	12	
12.	Is there a mark of terminal punctuation i+n?	17	19	
13.	Does i+2 have the code C (or K, L, N or O) ?	18	14	
14.	Does i+3 have the code D (or N or O)?	18	15	
15.	Does i+4 have the code E (or N) ?	18	19	
16.	Place a sentence separation boundary at i+n+m.	-	-	18
17.	Place a sentence separation boundary at i+n.	-	-	18
18.	Place a sentence separation boundary at i.	-		19
19.	Let i+1 be i.	-	-	A

The sentence separation routine has been designated for generalization and expansion. Accurate delimitation of machine sentences is basic to machine translation. It can prevent mismatchings across sentence boundaries. It can economize on machine time by setting definite limitations to the area within which structural cues are sought. Moreover, the structure of a machine sentence reveals its structural value in the composition of the text sentence which contains it, and leads to better integration of the analysis of the entire text sentence.

A description of a more highly generalized and more thoroughly exploited sentence separator operation can be found in the paper on English-to-Turkish Translation Research under the heading Segmentation.

IDIOMS AND COLLOCATIONS

The idiom operation is the first of the translation operations in the Se system, but in the Direct Conversion Programming it follows the exclusion operation. The change was made because the exclusion routine led to certain disarrangements of the idiom routine. Two solutions were possible: either modifying each of the routines, or changing their order. The second solution was adopted.

Definition of an Idiom

An idiom is here defined as a stretch of two or more words which cannot be translated by the general transfer system in its present form.

Not all such stretches are treated as formal idioms, however. If a string results in an unacceptable but more or less comprehensible output, and can clearly be subjected to generalized treatment, it is left as it stands in the expectation that the normal program of development will eliminate the difficulty.
those stretches which result in a completely garbled output are treated formally as idioms. In such cases, it is necessary to supply a gloss which will replace the glosses of the individual words which compose the idioms.

Some of these formal idioms will probably remain as idioms permanently. Others will be structurally analyzed and then fitted into one of the generalized patterns as further research is completed and programmed.

Studies are being conducted to determine what criteria make it advisable to assign a stretch to the idiom class. A pattern of periodic reviews of the idiom list has been established to insure that non-permanent idioms are removed from the list as quickly as is practical.

In the GAT, the translation of idioms is intended to be effected by the machine addition of numerical codes. This process is as anomalous to the GAT system as an idiom is felt to be to the system of the language in which it occurs but it continues to be used, since the expectation is that the number of idioms will never be large.

Coding of Idioms

The words which occur in any stretch formally designated as an idiom (idiom candidates) are entered in the unsplit dictionary, and receive the idiom diacritic. (The idiom diacritic is the digit 1 in position I:49. Since every item which has the idiom diacritic also has an idiom candidate code, the idiom diacritic is superrogatory, and is eliminated in some of those forms of the dictiary which are developed from the basic dictionary for specific purposes.)

Idiom candidates are divided into two classes. There are those which begin an idiom; these are called initials. There are those which do not begin an idiom; these are called sequents. The same word may occur both as an initial in one idiom and as a sequent in another. Each item which can occur as an initial has an identification code in positions I:50-53. Each item which can occur as a sequent has an identification code in positions I:54-57.

The Translation of Idioms

The English translations of the idioms are listed in an idiom dictionary, and this dictionary is keyed to the sum of the identification numbers of the initial and of the sequents of that idiom.

Consider an example. The following stretch was treated as an idiom. The initial and the one sequent are given with their idiom code numbers.

BROMISTOGO	VODORODA
15	24

The sum is 39. This gives the cue to the English translation. A search is made for the number 39 in the idiom dictionary and the following information is found.

39 (14 + 24): of hydrogen bromide

The translation is printed out as 'of hydrogen bromide'. Occasionally,

identical sums are composed of different components.

189_1 (16 + 127): phosphorous oxybromide

189_2 (123 + 66): to add to

189_3 (180 + 9) : aluminum chloride

There is no danger of ambiguity, however, as the operation checks for the components as well as for the total. In the dictionary the identical totals are distinguished with subscripts as shown in the above examples.

In cases where the idiom may be interrupted by an adjective which is not an essential part of the idiom, the code Y in position I:57 of the initial item of the idiom indicates this fact. The code X in the same position indicates that no such interruption is to be expected.

In cases where the sum of the identification codes of the initial and all of the sequents is not to be found in the dictionary, the identification code of the last sequent is subtracted from the total and a new attempt at matching in the dictionary is made.

When the idiom operation is completed, the assigned codes and English equivalents are stored in a suitably addressed location, and may be found as needed for other routines or for the final translation. The idiom routine can be generally outlined as follows:

The Procedure

Each textword in the sentence is analyzed. When the textword being analyzed (i) carries an idiom diacritic, the following tests are made.

		Y	N
1.	Does i have an idiom number as an initial (I:50-53) ?	2	18
2.	Register that number.	-	-
3.	Let a be 1 in what follows.	-	-
4.	Does i+a have an idiom number as a sequent (I:54-57) ?	5	7 -1
5.	Register that number.	-	- -
6.	Increase a by 1	-	-
7.	Does i have the code Y in I:57?	8	9
8.	Is i+a an adjective?	6	9
9.	Add all the numbers registered.		
10.	Is this sum matched by a number in the idiom dictionary?	11	15
11.	Does the matching number in the idiom dictionary have a subscript?	12	16
12.	Compare the numbers registered with the components of that idiom dictionary number.		
13.	Do the components match the numbers registered?	16	14
14.	Is there the same number with a larger subscript in the idiom dictionary?	12	15
15.	Subtract the last of the registered numbers from the sum.		
16.	Copy out the meaning of the idiom from the idiom dictionary and store it.	-	
17.	Let the item to the right of the idiom stretch be i.	-	
18.	Let i + 1 be i.	-	

Some stretches which are now listed as idioms are of a special class which the name 'collocation' has been assigned. These are stretches of two or more items which function as one structural unit. Thus V TECENIE is written as two items, but functions as a preposition; JELEZNA4 DOROGA is also written as two items, but functions like a simple noun.

In the case of idioms, it is only necessary to supply a gloss to replace the glosses of the words which compose the idiom. In the case of collocation it is necessary to supply both a gloss and structural coding to replace the glosses and codings of the items which compose the collocation.

The SLC system of programming collocates items in translating from French to English by means of local operations (operations dealing with a specific word rather than with a structural unit). The key word of the collocation, usually the least frequent, receives a code which calls on the local operation. This operation then searches the environment for the other items of the collocation in their appropriate positions, zeroes the translation of the other items, and substitutes a gloss and coding of the entire collocation for the gloss and coding of the key word. Presumably this method could also be used in Russian-to-English translation.

The Direct Conversion system of programming, which is used only for Russian-to-English, has not yet been called upon to collocate items, and so no procedure has been tested.

It would be of particular advantage to be able to recognize collocations during the lookup procedure. Ways of doing this have been suggested, but none seems to be both useful and compact.

EXCLUSION

When the text to be translated is highly technical, certain easily defined stretches of the input text can be transferred as they stand into the output language or can be translated item by item. Such stretches may therefore be excluded from the normal translation procedures.

Stretches which are candidates for exclusion are those which consist of abbreviations, formulae, numbers, the percent sign, the degree sign, punctuation marks, arithmetical signs, Greek letters, Roman numerals, anything written in the Latin alphabet, single words contained within formulae, and certain Russian words which can be defined by list. (Extensive research shows that, apart from single words in formulae and tables, an exclusion stretch el a minimum of two items. A one-item stretch usually has a function in the sentence, and its exclusion may seriously affect the structural analysis of the sentence.)

The exclusion routine is designed to recognize such stretches and to effect their transfer.

The exclusion operation is used when two conditions apply. First, the nature of the stretch must be such that it can be directly transferred. Secondly, the boundaries of the stretch must be able to be recognized by the machine. In order to facilitate recognition, the candidates for the exclusion stretch and the candidates for the exclusion boundaries carry certain identification codes.
Some of these are permanent in the dictionary entry, and some are automatically generated by the computer. When any of these codes is sensed, the exclusion operation extracts the exclusion stretch. The remainder of the sentence then goes through its normal analysis.

The words which are to have exclusion candidate codes are found by an. inspection of Russian texts and their English translation. A listing is made of the items which constitute the boundaries of each excludable stretch (boundary candidates) and of the items which constitute the excludable stretch itself (exclusion candidates). The items which are boundary candidates are not included in the exclusion stretch. Occasionally an item which occurs within an exclusion stretch is a boundary candidate elsewhere (e.g. , the sentence period.). In order to prevent the loss of information about boundaries in such a case, exclusion is effected when an item within the exclusion stretch can have a functional role in the non-excluded part of the sentence.

Because the same item may have both functions, the boundary candidates are coded separately from the exclusion candidates. The boundary candidate codes are in position I:47; the exclusion candidate codes are in position I:48.

The boundary candidates are permanently coded as follows.

L left boundary.
R right boundary.
A either left or right boundary.
Ø the item cannot be a boundary.

The exclusion candidates are coded as follows.

X exclusion candidate (permanent code).
Ø the item cannot be an exclusion candidate (permanent code).
E member of exclusion stretch (computer-generated code).

All boundary candidates and exclusion candidates are entered in the unsplit dictionary.

The exclusion operation checks every item in each sentence, and all items which have X in I:48 are registered. Any stretch with two or more items is treated as a potential exclusion area, and all exclusion candidates are included, unless the code R is sensed in I:47 of any item, in which case the item having that code is treated as the right boundary.

When an exclusion area has been identified, it is set aside for direct transfer with no rearrangement. In most cases the operation calls for a direct transliteration; Russian words and certain abbreviations are translated, but the translation is on a one-to-one basis with no insertion or lexical choice. When an abbreviation is also an idiom, its English equivalent is extracted from the idiom glossary by the idiom operation.

The following steps indicate the general nature of the exclusion operations.

	Y	N	A
1. Let the first word of the sentence be i.	-	-	2
2. Let a be Ø.	-	-	3
3. Does i+a have the code L in I:47?	6	4	-
4- Is i+a the last item in the sentence?	99	5	-
5. Increase a by 1.			3
6. Mark i+a as a left boundary (LB).	-	-	7
7. Let b be 1.			8
8. Does i+a+b have the code X in I:48?	9	10	-
9- Increase b by 1.	-	-	8
10. Is b equal to 2 or greater?	11	5	
11. Does i+a+b have the code R in I:47?	15	12	
12. Does i+a+b have the code L in I:47?	5	13	-
13. Does i+a+b +1 have the code X in I:48?	9	14	-
14. Decrease b by 1.	-	-	11
15. Is b equal to 3 or greater?	16	5	

16. Register all items from i+a+1 to i+a+b-1 inclusive as members of an exclusion stretch.

17. Register the exclusion stretch for direct transfer into the target language.
18. Let i+a+1 be i.

INTERPOLATION

When a stretch of numerals or formulas or symbols appears in a text, a speaker normally reads it off as a stretch of words. Many of these words, in a highly-inflected language such as Russian, are inflected words.

However, these symbols will appear in the output either in the same form as in the input (Arabic and Roman numerals, punctuation marks and symbols) or in a transliterated form (abbreviations, formulas). It would seem that the exclusion operation alone should handle them in a satisfactory manner. It does not. The most cursory examination of texts shows that the recognition of the inflections is essential to the structural analysis of the sentence, and that the same stretch of symbols may, in different situations, represent items with quite different inflections and therefore with quite different structural values.

If the structural effect of the inflection is confined within the limits of the symbol stretch, it need not be considered in the general translation. The exclusion routine removes the stretch and transfers it directly into the target language.

If the structural effect of the inflection extends beyond the limits of the symbol stretch, then the stretch acts as a part of speech and affects the form of some items outside the stretch, and is affected by the form of other items outside the stretch.

Thus, a stretch which represents an inflected form in Russian may require the insertion of a preposition or of the definite or indefinite article in English. This fact must be established, even though the stretch in its written form shows no suffix which could serve as a cue in the morphological analysis.

An operation has been developed to deal with such problems. This operation interpolates codes to indicate

a) what part of speech the symbol stretch represents,

b) what effect the symbol stretch has on the form of the items outside of itself, and

c) what effect the items outside the symbol stretch have on the stretch itself.

In cases where the symbol stretch contains a symbol whose value is ambiguous, an operation has also been developed which interpolates the codes necessary to resolve the ambiguity.

These operations are the interpolation operations.

As an example of some of the difficulties involved, the Arabic numerals may be cited.

The Arabic numeral 5 may correspond to the Russian forms P4T6 (nominative or accusative case), P4TI (genitive, dative or locative case), P4T6H (instrumental case). All of these are cardinals and correspond to the English 'five'. The Arabic numeral 5 may also correspond to P4TY1 (nominative or accusative), or any of ten other forms of the ordinal numeral. In this case it corresponds to the English 'fifth'.

As a cardinal, the numeral 5 will act as a noun and govern a noun in the genitive plural if it is in the nominative or accusative inanimate case, but will act as an adjective and agree with the noun if the sentence structure requires some other case. It is necessary, then, to determine how the sentence structure in general governs the symbol 5 before it is possible to determine how the symbol 5 governs or agrees with a noun.

Other considerations enter in. In translating V 5 OPYTAX, a human translator recognizes OPYTAX as plural, reconstructs the phrase as V P4TI OPYTAX, and so obtains the translation 'in five experiments'. In translating V 5 OPYTE, he recognizes OPYTE as singular, reconstructs it as V P4TOM OPYTE, and so obtains the translation 'in the fifth experiment'. (There are instances, especially in titles and footnotes, where a direct translation is possible. For example, GLAVA 5 can be translated 'Chapter 5' even though 5 corresponds to the Russian form P4TA4, that is, to a form of the ordinal, rather than to a form of the cardinal, numeral,)

It is the reconstruction of the text by a human translator which is replaced in machine translation by the interpolation routines.

The Scope of the Interpolation Operation

The linguistic items which may appear in a symbolic form and so require the use of interpolation routines are these:

a) numerals,

   i) Arabic (cardinal and ordinal),
   ii) Roman (ordinal),

b) abbreviations printed in Cyrillic letters,

c) formulas,

d) punctuation marks and symbols,

   i) colon,
   ii) parentheses,

iii) degree sign,
iv) percent sign,
v) equal sign.

Arabic Numerals Arabic numerals may function as nouns, adjectives and adverbs.

Nouns:

Numerals function as nouns if the numeral is part of a government structure which can be replaced in its entirety by a noun. Such numerals are all of those representing cardinal numbers which are in the nominative or accusative case, except for the numerals with Cue One. (Numerals are said to have Cue One if they both represent whole cardinal numbers and have the digit 1 in the units position, provided, however, that they do not also have the digit 1 in the tens position. Thus, 21, 131, 4651 have Cue 1, while 2, 6, 11, 17, 211, 23. 1 do not.)

Adjectives:

Numerals function as adjectives if the numeral is part of an agreement structure in which the numeral can be replaced by an adjective. These are:

a) the cardinal numerals which have Cue One, in whatever case they occur,

b) the cardinal numerals, other than those with Cue One, which are in any other case but the nominative or accusative,

c) the ordinal numerals in any case form, except those included under adverbs.

Adverbs:

Numerals function as adverbs if the numeral is part of a government or agreement structure which can be replaced in its entirety by an adverb. These are:

a) cardinal numerals in adverbial expressions such as time, proportion, distance.

b) cardinal numerals appearing as prefixes or suffixes to names of chemical compounds.

c) ordinal numerals which, in translation into English, appear as cardinal numerals in apposition to a noun (Chapter 5, Figure 3, etc The nouns which are followed in Russian by ordinal numerals but which are translated into English by apposition structures with cardinal numerals include the words STRANIQA; 'page', RISUNOJ 'figure', TABLIQA: 'table', GLAVA: 'chapter', CAST6: 'part', RAZDEL: 'section'.

The contexts in which Arabic numerals function as nouns, adjectives, adverbs were analyzed. This analysis has led to an operation which treats Arabic numeral on the basis of the class of the immediately preceding item (the determinant). Nine such subroutines have been established; the subroutine in which the determinant is a noun has been divided into six sections depending upon the case of the determinant.

These subroutines lead to the classification of each Arabic numeral as noun, adjective or adverb. It must be noted that, for machine translation purposes, numerals are sometimes classed as adverbs even though they are nouns or adjectives in standard grammatical terms.

Roman Numerals

All Roman numerals are excluded from the translation process. This is true whether they represent a Russian ordinal numeral which is translated in English by an ordinal numeral (V XX VEKE, that is, V DVADQATOM VEKE: 'in the twentieth century'), or a Russian ordinal numeral which is translated English by a cardinal numeral in an apposition structure (V VYPUSKE XX, that is, V VYPUSKE DVADQATOM: 'in Issue 20').

Abbreviations

Abbreviations function as nouns, verbs, adjectives, or adverbs, and are coded in the dictionary in a manner which indicates their function. But some abbreviations are ambiguously either noun or adjective. Interpolation is necessary to resolve this ambiguity in each specific environment. Again, those abbreviations which are nouns or adjectives occur in various case forms and these case forms are not indicated in the abbreviation. Interpolation is necessary to determine the case form in each specific environment.

Each abbreviation is entered in the unsplit dictionary. It is coded as follows in the first six positions of the parset location.

position	Code	Meaning
1-4	Ø 1 Ø Ø	Abbreviation
5	0	Part of a bound item
	1	Noun
	2	Verb
	3	Adjective
	4	Adverb
	5	Ambiguous (may be noun or adjective)
6	Ø	No idiom participation
	1	Idiom candidate

The English translation is given in the first English location. The only lexical choice involved in abbreviation transfer is in the case of idiom participation.

The interpolation routines for abbreviations follow this sequence:

a) Idiom participation is checked. See the paper on Idioms.

b) The ambiguity between the nominal and the adjectival function of the abbreviation is resolved. This is achieved by consideration of the form-class of the immediately following item.

c) The case of nominal and adjectival abbreviations is interpolated.

d) The abbreviation is translated.

Formulas

In the case of a formula, the interpolation routine depends primarily on the function category of the immediately preceding item, and secondarily on the function category of the immediately following item. There are a few cases where these are insufficient to determine the function category of the formula in any way; in such cases it is necessary to consider the lexical content of the environment. There are also pairs of cases where the determinants, though identical in both cases, surround a formula which is of one function category in one case, and of another function category in the other.

Such cases of ambiguity are quickly described.

The first possibility is that in which the immediately preceding item is a noun in the genitive case and the immediately following item is a reflexive verb.

The formula functions either as a noun in the genitive case or as a noun in the nominative case. Determining the function of the formula in the environment

entails the lexical identification of the immediately preceding item. If this is gram, millilitre, or some such measure, the formula is a noun in the genitive case. Otherwise, the formula is a noun in the nominative case.

The second possibility is that in which the preceding item and the following item may have any of a number of different functions, while the formula functions either as a noun in the genitive case (this calls for the preposition 'of' in the output) or as an adverb (this calls for no preposition in the output). Determining the function of the formula in this situation again entails the lexical identification of the immediately preceding item. The lexical items which determine that the formula is an adverb in this environment are defined by list (adverb determining list). Examples of items from the list are SOEDINENIE, TIP, SISTEMA, ZNACENIE, CLEN, PROIZVEDENIE, VELICINA, and FORMULA. If the immediately preceding item is a word on this list, the formula is an adverb. Otherwise, the formula is a noun in the genitive case.

Colon

Colon

A colon receives the coding 0603 in the first four positions of the parset. If the colon is not subject to exclusion, the following routine goes into effect. If the immediately preceding item and the immediately following item are both Arabic numerals or formulas, the colon is coded as a verb which governs the dative case. Otherwise, the colon is a mark of terminal punctuation.

Parentheses

If the exclusion operation does not exclude a parenthetical structure, then the parenthetical structure is subjected to the general translation process as unit separate from the rest of the sentence.

Symbols

The degree sign receives the coding Ø 2 Ø 2 in the first four positions of the parset. It is also coded as a noun. If the immediately preceding item is not adjective, the degree sign is in the genitive case. Otherwise, it is in the same case as the adjective.

The percent sign receives the coding Ø 2 Ø 1 in the first four positions of the parset. It is coded as an adjective either if the immediately preceding item is a formula or if the immediately following item is an adjective, a formula or a noun. Otherwise, the percent sign is coded as a noun. If the immediately pre-ceding item is an adjective, the percent sign is in the same case. If the immediately preceding item is a noun, the percent sign is in the genitive case.

The equal sign receives the coding Ø 7 Ø 4 in the first four positions of parset No interpolation operation is required for the equal sign itself. It is coded as

verb governing the dative. This coding is required only on a few occasions where the equal sign has not been excluded by the exclusion operation.

Order of Interpolation Routines

The first interpolation routine is that to determine the part of speech of the percent sign, wherever it occurs in the sentence. After that, the interpolation routines are used as required from left to right in the sentence.

GAP ANALYSIS

Sometimes the Russian textword is not found in the dictionary because of a misprint in the text, because of an error in the keypunching, or because of lacuna in the dictionary. The meaning of the Russian word is lost, and this loss in meaning cannot be bridged. The structural value of the Russian word is also lost, and this loss is more damaging since it may affect the analysis of the entire sentence. Fortunately, the structural gap can often be bridged sufficiently well that structural analysis of the sentence continues to be possible. The operation for bridging the structural gap is called Gap Analysis.

When a textword is not found in the dictionary, the computer generates code 8 in the parcue position (I:36), and gives the Russian word itself as its gloss. The result is, in effect, that the Russian word is printed out in the translation.

Even though the word has not been found in the dictionary, it is frequently possible to derive information about its structural function from its shape, and from its environment.

Each unrecognized word is compared with a list of characteristic endings of verbs, nouns, adjectives, and adverbs. These four parts of speech are chosen for two reasons. First, it is most probable that any word not in the dictionary will be one of these parts of speech; second, if it is not one of these parts of speech, the difficulty of recognizing its structural value from the shape of its ending will be insuperable.

The longer endings on the list, -when matched, are usually decisive in themselves. The shorter endings often require further substantiation from the environment of the unrecognized word. When enough evidence is acquired to make it seem reasonable that the unrecognized word is a particular part of speech, the computer generates a code in the parcue position to designate the part of speech. This code replaces the code 8 which was generated when the word was not found in the dictionary. Other codes are generated where possible and these help in the analysis of the remainder of the sentence. The unrecognized word is still printed out in its exact Russian form, of course, even though the computer may have acquired enough information to establish that it would normally be subject to the addition of endings during English synthesis.

Suppose the word MARGAETS4 occurs in the text as a result of a mis-punching when the text was converted to cards; the word will not be matched in the dictionary. The last letter is checked against the list of endings so as to narrow the selection down to those endings which terminate with -4. Then these endings are checked against the ending of the unrecognized form and a match is made with the ending -ETS4. The form is almost certainly an intransitive 1 verb in the third person singular of the non-past tense. The computer generates codes to convey this information.

Suppose the first item of a sentence is the word 5UR4 and the third item is a comma. The word 5UR4 is not matched in the dictionary because of a lacuna. The last letter of the unrecognized form is checked against the list of endings and a match is made with the ending -4. No match with a longer ending is possible.

A form ending in -4 may be the short form feminine of an adjective with so-called soft endings; the presence of a singular feminine nominal in the nominative and the absence of any finite verb except BYT6 are characteristics of the environment of such an adjective. If these conditions are fulfilled, 5UR4 will be recognized as an adjective.

Such a form may also be the present gerund of a verb; the characteristic position for a gerund is at the beginning of its segment, and a comma usually intervenes to the right before a finite verb occurs. If these conditions are fulfilled, 5UR4 will be recognized as a gerund.

Such a form may also be a number of different case forms of a noun, depending on the gender and the number of the putative noun; adjectives which agree with the noun in case or number, or an item which governs one of the possible case-forms are some of the charactersitics of the environment of a noun; if these conditions are fulfilled, 5UR4 will be recognized as a noun,

The tests are made in the order given on the principle of beginning with the possibility which is most easily tested and of proceeding to those possibilities which require more elaborate tests. In the case cited, 5UR4 will be recognized as a gerund, and the elaborate tests necessary for establishing it as a noun will not have to be made.

In such a manner as this, the gap analysis operation attempts to find evidence for the structural value of each unrecognized word so that the analysis of the entire sentence need not be crippled because of a gap in the structural information.

The gap analysis, as one may easily surmise, is a mass of detail, and it is not possible to give more than a general description here.

SYNTAGMATIC ANALYSIS

General

The morphological analysis treats of single words taken separately. syntagmatic analysis treats of groups, or stretches, of words linked by certain structural relationships.

Syntagmatic Stretches

I

The items which make up a syntagmatic stretch are described in two sets of terms. One set refers to the structural relationships among the members of the stretch; the other set refers to the positional relationships among the members of the stretch.

Structural relationships

Structurally, all stretches have at least two components. One is the head and the other is the dependent. The head is that structural element of the construction which can substitute for the whole construction in that context. Thus, in a Russian structure consisting of a verb governing a noun, the verb is the head, since the substitution of a verb only for the verb and noun combination acceptable, while the substitution of a noun only is not acceptable.

In some cases there is no structural element of the construction which i substitute for the whole construction; but there is one whose presence indicates that some different structural element can be regularly substituted for the whole construction. For example, in a Russian structure consisting of a preposition and a noun, neither preposition nor noun can regularly substitute for the whole, but it is a regular pattern that an adverb can substitute for the preposition and noun combination; this fact is conditioned by the presence of the preposition. The preposition may thus be regarded as the head of the construction.

The head and the dependent may each be expanded by the addition of other words; the words in these expansions may be associated in series, or linked either by punctuation or by a mediating word such as a conjunction or a participle; nonetheless the stretch, however, great the expansion of its component parts, is always analyzable as two syntagmatic elements: a head and a dependent.

Positional relationships

Positionally, since the words in a stretch follow each other in series, it is frequently useful to distinguish the first word (the initial of the stretch) from those that follow it (the sequents). It is also useful to distinguish the last word

(The final of the stretch) from the other sequents (the medials).

The coding of stretches

When the operation indicates that a series of items form a syntagmatic stretch, the computer generates a stretch code for each item in the stretch. The stretch code has six positions.

Positions 1 and 2 receive codes which indicate positional relationships; they vary depending on the position of the item in the stretch and on the positional interrelationship of different stretches; they do not vary with different structural types of stretch.

Positions 3, 4, 5 and 6 receive codes which indicate structural relationships; these codes do not vary from item to item in the stretch, but they vary with different structural types of stretch.

The first position receives the codes - (a hyphen) if the item is the initial of the stretch. It receives the code Ø (the digit zero) if the item is a medial in the stretch. It receives the code 1 (the digit one) if the item is the final of the stretch.

The second position receives no code unless there is a nesting. (See the paper on Nestings). If there is a nesting, the second position receives a code N for all items of all stretches in the nesting.

The third position receives a code indicating the part of speech of one, usually the prior, syntagmatic element in the stretch. In some stretches, this syntagmatic element is the head; in others it is the dependent.

The fourth position receives a code indicating the part of speech of the other, usually the latter, syntagmatic element in the stretch.

The fifth position receives a code which indicates the type of syntagmatic stretch. Homogeneous function stretches normally receive the code Ø, agreement stretches receive the code 1, government stretches receive the code 2, and modification stretches receive the code 3.

The sixth position receives a code which indicates the use of the declined forms in the stretch. If there is no declined form, this position receives other codes depending on the type of stretch for which the code is generated.

Since the positions from 3 through 6 receive the codes which are characteristic of the structural relationships of the items in the stretch, and since these Positions are unchanged for all members of the stretch, it has become usual to identify stretches by these four positions only. If, in the ensuing discussion, a stretch code is given as having four positions, it is understood that these are positions 3 through 6 of the six-position code.

Precise details of the coding in positions 3 through 6 will be given as a type of stretch is described.

Types of Syntagmatic Stretches

Syntagmatic stretches are of four types. These are homogeneous function stretches, modification stretches, agreement stretches, and government stretches. The GAT system treats them in that order, and they are described here in that order.

Homogeneous function stretches

When neighboring adjectival or nominal items have the same case, number and gender, there is a high probability that they have the same grammatical function. Such items are assigned a homogeneous function code by the computer

If all of the adjectives and the noun are unambiguous (that is, if they have a form which can only be of one case, one gender, and one number), each item in the stretch receives a stretch code of the following description.

The first two positions are standard for all stretches.

The third position receives the code 3, which indicates an adjectival.

The fourth position is zeroed in.

The fifth position receives the code Ø, which designates a homogeneous function stretch.

The sixth position receives a number from 1 to 6, which designates the particular declensional case common to the members of the stretch

If the adjectives or nouns are ambiguous in any way, (that is, if they have a form which can be of more than one case, more than one gender, or more than one number), each item in the stretch receives a stretch code in which positions 3 through 6 receive the letters HOMO. The stretch code HOMO is also used when two or more nouns share a homogeneous function in a particular case. I

The homogeneous function stretch codes are, in effect, pre-syntagmatic they bring about a preliminary linking of items of the same form and function and this shortens and simplifies the subsequent syntagmatic operation.

Modification stretches

Modification structures are combinations of an adverb as the dependent, with a verb, adjective, or another adverb as the head. Head and dependent form a structure characterized not by morphologically expressed case relation-ships, but by relative position.

The computer check for a modification relationship is a search for the part of speech which immediately precedes any adverb, and for the part of speech which immediately follows it. If no verb, adjective, (including participles), or adverb (including gerunds) is adjacent to the adverb being analyzed, the adverb being analyzed is disregarded in syntagmatic analysis and receives codes only in the subsequent syntactic operation. If a verb, adjective, or adverb is adjacent to the adverb being analyzed, the verb, adjective or adverb is linked with the adverb in a modification stretch.

The modification stretch receives a stretch code of the following description.

The first two positions are standard for all stretches.

The third position receives the code 4, which denotes an adverb.

The fourth position receives the code 2, if the head is a verb; the code 3, if the head is an adjective or a participle; and the code *4*, if the head is an adverb or a gerund.

The fifth position receives the code 3, which designates a modification stretch.

The sixth position receives the code P, if the head precedes the adverb being analyzed, or the code F, if the head follows it.

Types of modification stretch

There are three types of modification stretch.

1. The most frequently encountered modification stretch is the combination of an adverb with a verb or gerund.

NEREDKO	SLUCALOS6	'it frequently happened'
42 3F	423F	
VOOB5E	GOVOR4	'generally speaking'
423F	423F	

Here the verb SLUCALOS6: 'happened' and the gerund GOVOR4: 'speaking' are the heads, and the adverbs NEREDKO: 'frequently' and VOOB5E: 'generally' are the dependents.

2. The combination of an adverb with an adjective or with a participle also forms a modification stretch.

NAIBOLEE 433F	DREVNI1 433F	'most ancient'
DALEKO 433F	ZAWEDW11 433F	'far advanced'

Here the adverbs NAIBOLEE: 'most' and DALEKO: 'far' are the dependents, and the adjective DREVNI1: 'ancient' and the participle ZAWEDWI1: 'advanced' are the heads.

3. The combination of an adverb with another adverb also forms a modification stretch.

POCTI 433F	POSTO4NNO 433F	'almost constantly'

Here POSTO4NNO: 'constantly' is the head, and POCTI: 'almost' is the dependent.

Agreement stretches

Agreement stretches are combinations of an adjectival word as the dependent and a nominal word as the head; the adjectival is of the same gender, number, and case as the nominal word. (Nominal words include nouns, personal and reflexive pronouns, adjectives with noun function, and cardinal numbers in the nominative and accusative cases. Adjectival words are adjectives, participles, non-personal and non-reflexive pronouns, and ordinal numbers; the third person possessive pronoun forms, (EGO, EE, IX), which act in a manner similar to that of adjectives, but have no inflection such as that of adjectives, form a special group. These words are treated as adjectives and are assigned the same case as their head.)

Not all types of adjective-noun agreement which are possible in Russian are treated in syntagmatic analysis. Agreement between an item in the subject and an item in the predicate usually remains unanalyzed until the syntactic operation.

All members of an agreement stretch receive a six-position code of the following description.

The first two positions are standard for all stretch codes.

The third position receives the code 3, which signifies an adjectival.

The fourth position receives the code 1, which denotes a nominal.

The fifth position receives the code 1, which represents an agreement structure.

The sixth position receives a code which is a digit from 1 through 6; these digits designate the case in which the agreement structure occurs.

Types of agreement stretch

Unlike modification and government stretches, agreement stretches are not subdivided into types. One example is therefore sufficient.

IX	KATALITICESKOMU	DE1STVIH	'their catalytic action'
3113	3113	3113	

The adjectives IX:'their' and KATALITICESKOMU:'catalytic' as well as the noun DE1STVIH:'action' receive codes which designate an agreement structure in the dative case, even though IX: 'their' has no formal case marker of the dative case.

Although agreement stretches do not fall into a set number of patterns, they show certain characteristics which are of interest. A study of 1, 868 agreement stretches and the cases in which they occurred supplied the following information about the percentage of occurrences in each case and the range of the stretch to the left and to the right of the head.

	Percentage of Occurrences	Maximum range to Left	Maximum range to Right
Nominative	30. 1	no data	no data
Genitive	38. 2	5 words	5 words
Dative	3. 1	6	2
Accusative	9. 1	5	3
Instrumental	8.4	6	2
Locative	10.9	4	3

(There were isolated instances where the stretch extended to as many as eighteen words to the left and seventeen to the right because parenthetical constructions interrupted the stretch.)

In the majority of instances, the dependents of the agreement stretches consisted of one or two adjectives. There were twenty-eight cases of three adjectives in agreement with the head, and two cases of four adjectives in agreement with the head.

Government Stretches

Government stretches are of two types: strong and weak. 3

Strong government stretches are combinations of a nominal word as the dependent and another word, usually preceding it, as the head. The head governs the nominal word and determines its case; the head is here called a strong case determiner. Strong case determiners receive a coding which indicates the strong government relationship; the case determined in the nominal is coded in I:58-62.

Weak government stretches are combinations of a preposition as the dependent and another word, usually preceding it, as the head. The head determines the occurrence of the preposition, and the preposition in turn governs some nominal word and determines its case; the head is here called a weak case determiner. Weak case determiners receive coding which indicates the weak government relationship; the case determined in the nominal by the preposition is coded in I:63-67.

The computer analyzes a government structure by searching for the part of speech which follow the head or, rarely, which precede it. All members of a government stretch receive a stretch code of the following description.

The first two positions are standard for all stretch codes.

The third position receives the code 1 if the head is a noun, the code 2 if the head is a verb, the code 3 if the head is an adjective or a participle the code 4 if the head is an adverb or a gerund, and the code 5 if the head is a preposition.

The fourth position receives the code 1 to indicate that the governed element is a noun, the code 3 to indicate that it is an adjective with a noun function, or the code 5 to indicate that the head determines the occurrence of a preposition in a weak government structure. (The preposition is then analyzed separately as a strong case determiner.

The fifth position receives the code 2 to indicate that the syntagmatic stretch is a government structure.

The sixth position receives a code which is one of the digits from 2 to 6, to indicate the case in which the governed element occurs. The sixth position also receives the code Ø (zero) to indicate that the stretch is a weak government structure, or that the governed item an infinitive.

Types of government stretch

There are six types of government stretch.

1. The head is a verb. These stretches are either strong verb-government stretches (a), or weak verb-government stretches (b).

   a) Strong verb-government:

PROFESSOR	DAL	OPREDELENIE
	2124	2124
'The professor	gave	a definition. '

   b) Weak verb-government:

OBRA5AHT	NA	SEB4	VNIMANIE
2520	2520		2124
	5124	5124	
'attract	to	themselves	attention'

2. The head is a preposition.

POSLE	OPERAQII
5122	5122
'after	the operation'

3. The head is a noun.

ORGAN	ZRENI4
1122	1122
'organ	of sight'

4. The head is an adjective.

POLNI1	STRAXA
3122	3122
'full	of fear'

5. The head is a gerund.

DELA4	ZAMETKI
4124	4124
'making	notes'

6. The head is an adverb, especially a comparative adverb.

LUCWE	MEN4
4122	4122
'better	than I.'

The Overlapping of Syntagmatic Stretches

All three types of syntagmatic stretch overlap in the text.

V	DALEKO	ZAWEDWIX	SLUCA4X	GLAUKOMU
5126		5126	5126	
433F		433F		
		3116	3116	
			1122	1122
'In	far	advanced	cases	of glaucoma

Modification (433F) An adverb modifies a following adjective.

Agreement (3116) An adjective agrees with a following noun.

Government (5126) A preposition governs a stretch in the locative.

(1122) A noun governs a noun in the genitive.

Punctuation in Relation to Syntagmatic Stretches

Punctuation marks such as commas and dashes are most important in syntagmatic analysis. Commas and dashes may act as a boundary between machine sentences, or they may merely set off parenthetical constructions or items in a list, and thus not form such a boundary. The syntactic analysis requires a check of the words on either side of any comma or dash in order to determine whether the comma or dash is a machine sentence boundary. If it is, no apparent syntagmatic relationship that straddles such a boundary can be maintained. This is one of the chief reasons that the sentence separator open tion was moved forward from its former location in the syntactic analysis operation.

The Importance of the Syntagmatic Stretch

Since a syntagmatic stretch consisting of a head and dependents is replace-able by the head alone in most cases, the syntagmatic analysis in effect organ: extensive strings of words into structural units. These syntagmatic units are basic working material for the syntactic analysis. The syntagmatic units are

also the basis of the transpositions in the rearrangement operation. For example, a sentence is translated without rearrangement from the Russian as:

> 'Together with this in the reaction of oxidation became noticeable certain additional influences'.

Syntagmatic analysis produces stretch codes for the Russian items corresponding to together with this, in the reaction of oxidation, and certain additional influences, thereby organizing them into units and permitting them to be rearranged as blocks to produce a smoother English sentence:

> ' Together with this certain additional influences became noticeable in the reaction of oxidation'.

SYNTAGMATIC ANALYSIS

Nestings

The analysis of nestings is a special subdivision of the general syntagmatic analysis.

A nesting is a text sequence in which one syntagmatic stretch is introduced, but, before it continues to the final, another syntagmatic stretch is **being** introduced and concluded, and only then does the first syntagmatic stretch continue to its final. If the items of the interrupted stretch are designated by the number 1, and the items of the inserted stretch by the number 2, the form of the nesting may be schematized as

Initial 1 Initial 2 Final 2 Final 1.

It is characteristic of nestings that the points of division between the stretched are not marked by commas or by other punctuation.

Types of Nesting

Nestings fall into two groups. There are free nestings and bound nestings

A free nesting consists of a stretch which is interrupted by an independent stretch; the inserted stretch is independent in that its initial word is not the initial word of the interrupted stretch.

A bound nesting consists of a stretch which is interrupted by a dependent stretch; the inserted stretch is dependent in that its initial word is the same as the initial word of the interrupted stretch.

Nestings also fall into two groups which cut across the division of free and bound nestings. There are government nestings and agreement nestings.

A government nesting is one in which the interrupted stretch is a government stretch, and the inserted stretch is a government stretch.

An agreement nesting is one in which the interrupted stretch is an agreement stretch, and the inserted stretch is a government stretch.

From these definitions it follows that the inserted stretch is always a government stretch.

A free nesting is always a government nesting.

A bound nesting may be either a government nesting or an agreement nesting.

Free Nestings

A free nesting, then, is a sequence in which the initial of the interrupted stretch governs a nominal structure, but is separated from that nominal structure by one or more inserted structures. Each inserted structure also has an initial which governs a nominal structure. Thus, the order of a free nesting may be represented as follows. (The interrupted sequence is conventionally assigned the number 1.)

Initial 1--/Initial 2-Final 2/--Final 1

or, more briefly,

I:1--/I:2-F:2/--F:1.

Sequents other than the final are called medials (M), and they usually occur immediately before their respective final.

Examples:

OBRAZOVANIEM	I:1	-Ø1122			by formation
V	I:2			-Ø5126	in
PRIRODE	F:2			1Ø5126	nature
MINERALOV	F::l	1Ø1122			of minerals
UDAVALOS6:1				-Ø2220	it was possible
PRI	1:2	-Ø5126			with
SOOTVETSTVUH5EM	M:2	ØØ5126			appropriate
IZMENENII	M:2	1Ø5126/-	Ø1122		change
USLOVIIF:2			1Ø1122		of conditions
POLUCAT6F:l				1Ø222 Ø	to obtain

In this nesting, the initial of the interrupted stretch (1:1) is a finite verb and the final of the interrupted stretch (F:l) is an infinitive; one or more independent stretches nest between initial 1 and final 1. The relationship between the finite verb (1:1) and the infinitive (F:l) is neither that of government nor of agreement in the terminology of the GAT analysis. Accordingly, this type of nesting is not usually described as a free nesting, although it is treated as one.

Bound Nestings

Bound nestings are either agreement nestings (all agreement nestings are necessarily bound) or bound government nestings.

Agreement Nestings

An agreement nesting is a sequence in which the initial of the interrupted stretch (I:l&2) is an adjective or a participle. The initial 1&2 both is a member

of an agreement stretch in the interrupted stretch (I:1) and governs a nominal structure in the inserted stretch (I:2). The sequents of initial 1&2 as a **stretch** case determiner continue to their final (F:2) and then the sequents of the **initial** as the beginning of an agreement stretch continue to their final (F:l). Thus, the order of an agreement nesting may be represented as I:1&2--F:2/--F:1,

Examples:

BOGATYE	I:l&2	- Ø3111/ -Ø3125	rich
METANOM	F:2	1Ø3125	in methane
PRIRODNYE	M:l	ØØ3111	natural
GAZE	F:l	1Ø3111	gases
BOGATA4	I:l&2	-Ø3111/- Ø3125	rich
CO_2	F:2	1Ø3125	in CO_2
BEDNA4	M:1/I:3	ØØ3111/-Ø3 125	poor
KISLORODOM	F:3	1Ø3125	in oxygen
ATMOSFERA	F:l	1Ø3111	atmosphere

Bound Government Nestings

A bound government nesting is a sequence in which the initial of the interrupted stretch (I:l&2) is a noun or a verb. This initial 1&2 both governs one nominal structure in the inserted stretch and governs another nominal structure in the interrupted stretch. The inserted government stretch continues to its final (F:2) and then the interrupted government stretch continues to its final (F:l). Thus, the order of a bound government nesting may be represented in the same way as the order of a bound agreement nesting (I:1&2- -F2 / - -Fl).

Examples:

RAZLOJENIE	I:l&2	-Ø1122/- Ø1125	decomposition
VODO1	F:2	1Ø1125	by water
NITRIDOV	F:l	1Ø1122	of nitrides
SOOB5AET	I.:1&2	-Ø2124/- Ø2123	imparts
ORGANIZCESKO1	M:2	ØØ2123	to organic
MOLEKULE	F:2	1Ø2123	molecule .
XARAKTER	F:l	1Ø2124 nature	

Mixed Nestings

There may be a combination of bound and free nestings in one complex. An independent stretch may be inserted into a bound nesting so as to produce a free nesting contained within the bound nesting.

OPUBLIKOVANNYE	I:l&2	-Ø3124/-03125	published
DO	I:3	-Ø5122	until

NASTO45EGO	M:3			-Ø5122	present
VREMENI	F:3			1Ø5122	time
AVTORAMI	F:2		1Ø3125		by authors
SINTEZY	F:l	1Ø3124			syntheses

Recognition and Coding of Nestings

Nestings are recognized and coded to permit of rearranging the word order prior to translation.

Recognition

The first step in recognizing a nesting consists of determining the relative positions of the initials and finals of the stretch codes involved. Each sentence is scanned from left to right, until a candidate for a stretch initial is encountered.

Should an initial have more than one government code, a special cleaning operation eliminates those government codes which are found to be inapplicable in the particular context.

If, after the initial of a stretch has been encountered, its final is not encountered before the initial and final of another stretch have been encountered, then everything from the first initial to the last final is recognized as a nesting; this is a free nesting.

If an initial is encountered which calls for more than one final, and these finals are successively encountered, then everything from that initial to the last final is also recognized as a nesting; this is a bound nesting.

For the purpose of recognizing nestings, only those initials which govern nominals are considered, since even in an agreement nesting the initial must govern a nominal in the inserted stretch as well as taking part in the agreement stretch. But nouns or strings of nouns in the genitive case, if they depend on a noun which is in an inserted structure, are ignored in scanning for nestings, even though they do govern nominals.

Coding

Once a nesting has been recognized, the machine generates a code to mark that nesting. It will be remembered that the second position of the syntagmatic stretch code is still vacant; this position receives the code N for every item from initial 1 to final 1 inclusive, regardless of the stretch to which the item belongs. Consider the sentence:

DAET S KATIONAMI R4DA METALLOV TRUDNO RASTVORIMYE I XARAKTERNO OKRAWENNYE SOEDINENI4.

'It produces, with the cations of a series of metals, hardly soluble and characteristically colored compounds.'

This sentence would be coded syntagmatically as follows:

DAET	-Ø2124					
S		-Ø5125				.V
KATIONAMI		1Ø5125	-Ø1122			
R4DA			1Ø1122	-Ø1122		
METALLOV				1Ø1122		
TRUDNO					-Ø433F	
RASTVORIMYE					1Ø433F	-Ø31l1
XARAKTERNO					-Ø433F	
OKRAWENNYE	ØØ2124				1Ø433F	ØØ3111
SOEDINENI4	1Ø2124					1Ø3111

In scanning from left to right in the text (but from top to bottom in the above table) the initial DAET (-Ø2124) is encountered; then, before its final (1Ø214) is encountered, another initial, S (-Ø5125), is encountered. The presence of a nesting is recognized. The initial of the first inserted stretch is the word S. The final of the first inserted stretch is KATIONAMI (1Ø5125). But KATIONAMI is also the initial (-Ø1122) of another stretch of which R4DA is the final. However, KATIONAMI R4DA is not recognized as an inserted stretch, although KATIONAMI has the qualifications of governing a nominal and of being the initial of a stretch; such genitive case government by a noun in one inserted stretch (S KATIONAMI) is not treated in the GAT as requiring the recognition of another inserted stretch. Instead, R4DA is assigned to the same stretch as KATIONAMI; the first inserted stretch becomes S KATIONAMI R4DA. In exactly the same way, R4DA METALLOV is not recognized as a new inserted stretch, but METALLOV is amalgamated with KATIONAMI R4DA, and, consequently, with the inserted stretch. The first inserted stretch now becomes S KATIONAMI R4DA METALLOV. While TRUDNO has one qualification as the initial of an inserted stretch in that it is an initial (-0433F), it lacks the other qualification in that it is not a strong case determiner.

Again, RASTVORIMYE has one qualification, in that it is an initial (-Ø111 and might have the second qualification in that, as a passive form, it governs the instrumental case. But there is no instrumental case form except KATIONAMI and that is already assigned to the government of S; the assignment is made because of the higher priority enjoyed by S (a preposition) than by RASTVORIMYE (a participle), and because of position (the preposition immediately precedes its object). Therefore, RASTVORIMYE has no form to govern and is not a strong case determiner in this sequence; it is disregarded as the initial of an inserted stretch. The remarks applied to TRUDNO apply equally to XARAKTERNO.

Thus, the sentence is analyzed as a nesting, involving one inserted structure.

DAET / S KATIONAMI R4DA METALLOV / TRUDNO RASTVORIMYE I XARAKTERNO OKRAWENNYE SOEDINENI4.

All of the stretch codes associated with any of the words from DAET to SOEDINENI4 inclusive receive the code N in the second position of the machine word.

DAET	I:1	-N2124				
S	I:2		-N5125			
KATIONAMI	M:2		1N5125/-N1122			
R4DA	M:2			IN1122/-N1122		
METALLOV	M:2				1N1122	-N433F
TRUDNO	M:l					-N433F/-N3 111
RASTVORIMYE	M:1					
I	M:1					
XARAKTERNO	M:1					-N433F
OKRAWENNYE	M: 1	0N2124				1N433F/ØN3 111
SOEDINENI4	F: 1	1N2124				1N3111

In certain types of bound nestings in which initial 1 governs or agrees with the genitive case, a succession of nouns in the genitive case may produce ambiguity and so require special attention. In the sequence SOSTAV POLUCENNOGO GIDRILIZOM (a formula) GIDROZOLA OKISI JELEZA, initial 1 (SOSTAV) might govern any of the underlined nouns, which are all in the genitive case. Only semantic criteria can induce the acceptable analysis, which is as follows:

SOSTAV	I:1-N1122			composition
POLUCENNOGO	M: 1/I: 2	ØN1122/-N3125		obtained
GIDRILIZOM	F:2/I:3		1N3125/-N1122	by hydrolysis
(a formula)	F: 3		1N1122	of (formula)
GIDROZOLA	F: 1/I:4	1N1 122/-Nl 122		of hydrosol
OKISI	F:4/I:5		1N1 122/-N1122	of oxide
JELEZA	F:5		1N1122	of iron

If initial 1 governed another case than the genitive, there would be no ambiguity. The example S PEREMENNYM V ZAVISIMOSTI OT PRIRODY KATIONA SODERJANIEM is to be analyzed as follows:

S	I:1	-N5125				with
PEREMENNYM	M: 1/I:2	0N5125	-N2520			varying
V	F:2/I:3		1N2520	-N5126		in
ZAVISIMOSTI	F:3/I:4			1N5126	-N1520	dependence
OT	F:4/I:5		-N5122		1N1520	on
PRIRODY	F:5/l:6		1N5122	-N1122		nature
KATIONA	F:6			1N1122		of cation
SODERJANIEM	F:1	1N5125				content

If a compound preposition such as V TOKE or V TECENIE is treated as two distinct items, difficulties occur; if, however, compound prepositions are treated as collocations, and therefore as single items, the difficulties can be avoided.

In the example

PRIKOSNOVENII S VOZDUXOM PROKALENNOGO V TOKE VODORODA TITANA

concerning the contact with air of titanium which has been annealed in a current of hydrogen',

using the present procedure, the word VODORODA is recognized as final 1 (PRIKOSNOVENII being initial 1). This is not acceptable; it is the succeeding item TITANA that is final 1. However, if the phrase V TOKE is construed as a compound preposition (initial 3) governing a nominal in the genitive case, the present procedure is successful, since, in that case, VODORODA is recognized as final 3. Then TITANA is recognized as final 1, and the translation is acceptable.

Weak government is coded by assigning the same government code number (with zero as the last digit) to the head and to the preposition, and, in addition, the usual government code number to the preposition and to the word governed by it. This can be shown in an example:

IZ	-N5122			from
NERASTVORIMYX	0N5122	-N3520		indissoluble
V		1N3520	-N5126	in
VODE			1N5126	water
SILIKATOV	1N5122			silicates

The participle NERASTVORIMYX determines the occurrence of the preposition V (3520) which in turn governs the noun VODE (5126).

Rearrangement of Nested Structures

A nesting must often be rearranged before it can be translated acceptably into English. Government nestings are rearranged differently from agreement nestings.

A. In a government nesting, bound or free, everything which follows the last inserted stretch is shifted to a point immediately after initial 1. The words shifted may be

   a) final 1 (this may consist of two or more nouns in the same case, connected by a conjunction);
   b) all the nouns in the genitive case that follow final 1, the first of these nouns being governed by final 1 and each subsequent noun being governed
      by the noun which precedes it;
   c) adjectives that qualify any of the above nouns;
   d) adverbs that qualify any adjectives under (c);
   e) conjunctions that connect any of the above words;

f) particles which occur between any of the above words.

Thus, the nesting

OT	/VZVEWENNYX	VYWE/	TRUBCATYX	3LEKTROFIL6TROV
from	suspended	above	tubular	electro-filters

is rearranged to

OT	TRUBCATYX	3LEKTROFIL6TROV,	VZVEWENNYX	VYWE.
from	tubular	electro-filters	suspended	above.

Similarly,

NAGREVANIEM	/DO	300	GRADUSOV/	SMESI	AMAL6GAMY
heating	to	300	degrees	of a mixture	of an amalgam

OLOVA
of tin

is rearranged to

NAGREVANIEM	SMESI	AMAL6GAMY	OLOVA	DO	300	
heating	of a mixture	of an amalgam	of tin	to	300	

GRADUSOV.
degrees.

There are two modifications of the rearrangement procedure for government nestings.

1. If a multiple nesting contains participles in two or more inserted stretches, and if these participles each govern a nominal word in strong or in weak government, then, in addition to the re-arrangement prescribed in the general procedure, the word 'and' must be inserted between any two contiguous inserted stretches that contain participles.

Thus, the nesting

IZ	/POLUCAEMYX	ISKUSSTVENNYM	PUTEM/	/NERASTVORIMYX	
from	obtained	by synthetic	methods	indissoluble	

V	VODE/	SILIKATOV
in	water	silicates

is rearranged to

IZ SILIKATOV POLUCAEMYX ISKUSSTVENNYM PUTEM
from silicates obtained by synthetic methods and

NERASTVORIMYX V VODE.
indissoluble in water.

2. If there is one participle in the nesting which governs a nominal either strongly or weakly, and if that participle is in the last inserted stretch, then the segment to be shifted is shifted to a point immediately preceding the last inserted stretch instead of immediately preceding the first inserted stretch, and the word 'and' is not inserted.

Thus, the nesting

DE1STVII /NA NIX/ /LEGKO OTDAH5EGO SERU/
of action on them slightly giving off sulphur

POLISUL6FIDA AMMONI4

of polysulphide of ammonia

is rearranged to

DE1STVII /NA NIX/ POLISUL6FIDA AMMONI4 /LEGKO
of action on them of polysulphide of ammonia slightly

OTDAH5EGO SERU/.
giving off sulphur.

B. In an agreement nesting, everything which follows the last inserted stretch, is shifted to a point immediately before initial 1.

Thus, the nesting

OBRAZUH5IES4 /PRI VZA1MODE1STVII DVUOKISEl ILI IX
which form upon interaction of dioxides or their

GIDRATOV S SIL6NYMI 5ELOCAMI/ SOLI
hydrates with strong alkalis salts

is rearranged to

SOLI, OBRAZUH5IES4 PRI VZA1MODELSTVII DVUOKISE1 ILI
salts, which form upon interaction of dioxides or

IX GIDRATOV S SIL6NYMI 5ELOCAMI.
their hydrates with strong alkalis .

If final 1 of an agreement stretch is preceded by the final of a government stretch (final 2) which is also associated with initial 1&2 but is separated from it by other inserted stretches, then everything which follows the last inserted stretch is shifted to a point immediately before initial 1, and, in addition, the member of the government stretch and its qualifiers are shifted to a position immediately following initial 1.

Thus, the nesting

I:l&2	I:3	M:3	F:3	F:2
OPUBLIKOVANNYE	/DO	NASTO45EGO	VREMENI/	/AVTORAMI/
published	up to	the present	time	by authors

F:l
SINTEZY
syntheses

is rearranged to

SINTEZY, OPUBLIKOVANNYE AVTORAMI DO NASTO45EGO
syntheses, published by authors up to the present

VREMENI.
time.

Cases such as this are very infrequent.

Frequency of Nestings in the Russian Language

This analysis of nestings was based on the examples found in 50, 000 words of running text in the field of Organic Chemistry. A total of 551 nestings was found.

The nestings varied in length from three to thirteen words; the average length of a nesting was 5. 3 words. There was one nesting in each 90. 7 words of continuous text.

Initial 1 was a	noun	in 223 cases,
	participle	in 118 cases,
	verb	in 85 cases,
	preposition	in 67 cases,
	adjective	in 57 cases,
	adverb	in 0 cases.

Because of the relatively small number of nestings which served as examples, it is not possible to feel that these procedures for dealing with nestings are definitive. The procedures described in this paper are interim procedures only. This operation has nonetheless been programmed.

Operations

The syntagmatic operations are so arranged that a description of each operation in the order in which it is applied would involve much repetition if the nature of the operations is to be clear.

The syntagmatic operations are here described in the order in which they are applied. The preceding discussions of the general nature of the syntagmatic analysis and of nestings should help the reader understand the trend of the operations more easily than if each were discussed only in the sequence in which it is applied in the GAT.

The Preposition Cleaning Operation

Most Russian prepositions present no structural ambiguity. They are unique in form and indeclinable. They govern a nominal structure in a specific case, and the nominal structure almost invariably follows the preposition. A certain number of prepositions are ambiguous, however, in that items having the same shape govern different cases in the nominal structure and are associated with different meanings depending on the case which they govern. Thus, the Russian word S is always a preposition and can be roughly glossed as 'from' if it governs the genitive case, or as 'about the size of' if it governs the accusative case, or as 'with' if it governs the instrumental case.

The operation of the syntagmatic analysis which resolves such ambiguities in prepositions is called the Preposition Cleaning Operation. Each occurrence of an ambiguous preposition is examined in terms of the immediately following nominal structure. If the nominal structure is unambiguous as to case, the ambiguity of the preposition is immediately resolved. If the nominal structure is also ambiguous, a comparison of the array of cases governed by the preposition with the array of cases represented by the nominal structure shows which cases are common to both. Such a comparison usually resolves the ambiguity. Failing a complete resolution, at least some reduction of the ambiguity can be expected. If the combination of preposition and nominal structure remains ambiguous after comparison, a resolution is effected by the government operator or by the lexical choice operation.

For example, S is coded as governing the genitive (2) accusative (4) am instrumental (5) cases. This coding is entered in the dictionary as 20450 in I:58-62. (This government code has a position for all cases except the nomina-tive on the assumption that the nominative case is never governed; the coding therefore, occupies five positions.)

The noun-form USPEXOM is coded as instrumental (5) singular; (this coding has been generated as 000050 for the singular in III:66-71, and as ØØØØ

for the plural in III:72-77). In the phrase S USPEXOM: 'with success', a comparison shows the following:

S	20450
USPEXOM	000050
	000000

Since the only case common to the codings of these words is the instrumental (5), the ambiguity of the preposition is resolved immediately. Therefore, S is here a determiner of the instrumental case.

The adjective-form BOL6WIM is coded as instrumental (5) singular and as dative (3) plural. In the phrase S BOL6WIM USPEXOM: 'with great success', a comparison shows the following:

S	20450
BOL6WIM	000050
	003000
USPEXOM	000050
	000000

The ambiguity of the preposition is resolved by comparison with the ambiguous adjective BOL6WIM because the instrumental (5) is the only feature the two ambiguous forms have in common.

The following noun USPEXOM confirms the resolution, since USPEXOM, as the first compatible noun or pronoun following the preposition, is the head of the nominal structure governed by the preposition.

The noun RADOST6H is coded as instrumental singular; the same remarks apply to it in the phrase S RADOST6H: 'with joy' as apply to USPEXOM in the phrase S USPEXOM.

The adjective BOL6WO1 is coded as nominative (1), genitive (2) dative (3), accusative (4), instrumental (5) and locative (6) singular. In the phrase S BOL6WO1 RADOST6H: 'with great joy' a comparison shows the following:

S	20450
BOL6WO1	123456
	000000
RADOST6H	000050
	000000

The ambiguity of the preposition is neither resolved nor reduced by comparison with BOL6WO1 because the adjective is ambiguous in every feature in which the preposition is ambiguous. But all ambiguities are again resolved by comparison with RADOST6H, which is unambiguously instrumental.

The noun RADOSTI is coded as genitive dative and locative singular; if it occurs as a plural in any text, it will also be coded as nominative and accusative plural.

In the phrase S BOL6WO1 RADOSTI: 'as a result of (his) great joy', a comparison of codings shows the following:

S	20450
BOL6WO1	123456
	000000
RADOSTI	023006
	100400

The genitive case (2) is common to all three items, and the accusative case is also common. A check on the gender of RADOSTI (feminine) eliminates the possibilities of nominative (1) and accusative (4) in BOL6WO1, since BOL6WO1 has these values only when it is masculine. Or a check on the number of the adjective and of the noun eliminates the accusative case (4), since BOL6WO11 only accusative in the singular, and RADOSTI only in the plural. Either check eliminates the accusative (4). Therefore, S is a determiner of the genitive case

There are remarkably few cases in which a preposition continues to be ambiguous after comparison with the following nominal structure. In those few cases, other types of information help resolve ambiguity. In the case of S, 1 there is the fact that S with the accusative is obsolescent to the degree that is possible to list exhaustively all of the objects which it governs. If none of the objects if found, the preposition almost certainly does not govern the accusative (This type of resolution is made by the lexical choice operation,) There is also the fact that S with the genitive case is extremely unlikely to have an animate noun as its object unless it follows one of a certain of verbs, usually those that are compounded with the prefix S-. (This type of resolution is also made by lexical choice operation.)

The Nesting Operation

The nesting operation searches out and codes nestings. The general nature of this operation is described in the paper on nestings.

The Adjective Cleaning Operation

The adjective cleaning operation is similar to the preposition cleaning operation.

Adjectives which are ambiguous in number and case are compared with the contiguous nouns and adjectives in the expectation that the comparison will reduce the ambiguity, if not resolve it. The codes generated (or entered) in III:66-77 are compared in each case; those which are common are kept, and

those which differ are rejected.

For example: BOL6WO1 CELOVEK: 'large person'

BOL6WO1	123456 000000
CELOVEK	100000 020000

Therefore, BOL6WO1 is nominative singular. The ambiguity is resolved.

BOL6WO1 STRANE: 'large country'

BOL6WO1	123456 000000
STRANE	003006 000000

Therefore, BOL6WO1 is dative or locative singular. The ambiguity is reduced. (Since the locative case occurs only after certain prepositions, a search for a preposition to the left will resolve the ambiguity. If there is a preposition which governs the locative, BOL6WO1 is locative. If there is a preposition which governs the dative, or if there is no preposition, BOL6WO1 is dative. If there is the preposition PO, which governs the dative or the locative, a check of the environment for the object of the preposition or for certain verbs will give evidence as to whether BOL6WO1 is dative or locative in this case, with the strong possibility that it is dative, because the use of PO with the locative is obsolescent in Russian.)

BOL6WO1 RUSSKO1: 'large Russian woman'

BOL6WO1	123456 000000
RUSSKO1	023056 000000

Therefore, BOL6WO1 is genitive, dative, instrumental or locative singular. The ambiguity is reduced. A larger environment is necessary before resolution is possible.

The Noun Cleaning Operation

The noun cleaning operation is similar to the adjective cleaning operation.

Nouns which are ambiguous in number and case are both compared with the contiguous adjectives and checked against the items in the environment which govern nouns in specific cases; the expectation is that the comparison will reduce or resolve the ambiguity. The codes generated (or entered) in III:66-77 for the noun-form are compared with the codes generated or entered in those positions for the adjective-form. These same codes in the noun-forms are compared with the government codes of each case-determining item in the vicinity. The government codes are located in positions I:58-62; there are only five positions, since no case determiner governs a noun in the nominative.

For example: 3TOM POLE: 'this field'

3TOM	000006 000000
POLE	100406 000000

Therefore, POLE is locative singular. The ambiguity is resolved.

IZ 3TO1 CASTI: 'from that part'

CASTI	023006 100400
3TO1	023056 000000

Therefore, CASTI is genitive, dative or locative singular. (3TO1 is also in one of these cases, as will already have been determined in the adjective cleaning operation.) The ambiguity is reduced.

IZ governs 20000

Therefore, CASTI is genitive singular, and consequently 3TO1 is genitive singular. The ambiguity is resolved.

NE KUPIL KNIGI: 'did not buy the book'

KNIGI	020000 100400
KUPIL	governs 00400

On this basis, KNIGI is recognized as accusative plural, and this is probably wrong. All verbs which govern the accusative case must be checked for negation when they are negative, and in formal style, such verbs govern the genitive case instead of the accusative. Therefore, because of the presence of the particle NE, KNIGI is recognized as genitive singular, and the ambiguity is resolved.

Agreement Operation

The agreement operation establishes the agreement stretches and assigns agreement stretch codes as described in the paper on syntagmatic analysis.

All of the operations from the preposition cleaning to the agreement operation are brought together into one routine in the Direct Conversion Programming. This routine is called Syntagmatic I.

Government Operation

The government operation establishes the government stretches and assigns government stretch codes as described in the paper on syntagmatic analysis. This operation is a separate routine of the Direct Conversion Programming and is called Syntagmatic II.

SYNTAGMA TIC ANALYSIS

Further Research: Noun-Noun Structures

At present, the GAT syntagmatic analysis, though satisfactory in many ways, exhibits a pattern of shortcomings which prompts the suggestion that further coding of parts of speech may prove useful.

One of the shortcomings of the translation lies in its treatment of noun-noun structures, that is, of a noun in the genitive (N2) following another noun (Nl). Under the present syntagmatic rules, both nouns are coded as members of a government stretch (1122). The relationship of the initial noun to some other word in the sentence is coded in a diversity of ways, of course, depending on the structure of the sentence. One of the most frequent ways is the genitive government stretch (X12X), including the noun-noun government stretch (1122). If a genitive noun is the first noun in a noun-noun government stretch, it is necessarily governed by some form other than a noun, since otherwise the governing noun would also be included in the stretch, and the noun in the genitive could not be the first noun. If a genitive noun is not the first noun in a government stretch, then, excluding cases where the nouns are linked by a conjunction or are in apposition, the following noun is necessarily governed by the preceding noun.

These facts have given rise to the present routine for genitive noun government structures. The routine may be outlined as follows:

Noun in Genitive Case Routine:	Y	N	A
1. Is the noun in the genitive case the first noun in the government structure?	2	3	-
2. Transfer the genitive case ending as zero.	-	-	99
3. Transfer the genitive case ending by means of the preposition 'of'.	-	-	99

This routine has more detail than is given here, but the details do not alter the basic pattern.

Even in its more elaborate form, this routine is not able to handle certain difficulties in translation.

There are two types of difficulty which arise from this routine. One type of difficulty is caused by the assumption that any noun in the genitive is governed by any immediately preceding noun, whatever its case may be. This is the syntagmatic difficulty. The other type of difficulty is caused by the assumption that the English preposition 'of is an adequate transfer of the genitive case morpheme wherever a genitive noun follows another noun. This is the transfer difficulty.

The genitive noun government routine is based on the assumption that any noun in the genitive is governed by any immediately preceding noun, whatever its case may be. This assumption is not always valid. Here are two examples where it is not.

1. The noun which seems to be the initial (N1) belongs to an inserted structure in a nesting, while the noun which seems to be the sequent (N2) belongs to the interrupted stretch.

VSE	SKOPIVWIES4	ZA	DEN6	TUCKI	
3111	3111			3111	acceptable
		5124	5124		acceptable
			1122	1122	not acceptable
			N1	N2	

'All the small clouds which gathered during the day'.

2. The noun which seems to be the sequent (N2) is governed by the predicate of the sentence.

DURAKI	NANESLI	LESU	U5ERBA	NE	MEN6WE	X15NIKOV	
	2122		2122				acceptable
	2123	2123					acceptable
		1122	1122				not acceptable
		N1	N2				

'Vandals have done no less harm to the forests than commercial exploiters'.

These examples allow of various solutions. One is that the syntactic structure is to be established before the syntagmatic. This would involve a complete reorientation of the GAT. Another is that a hierarchy of government structures must be established. This seems to be the more satisfactory and practical solution at this point.

The basis for a hierarchical treatment of genitive government structures can be illustrated by samples of the translation of the genitive case. In some cases, the transfer is not very involved.

RAZRUWENIE CASTI PROIZVODITEL6NYX SIL
'The destruction of a part of the productive forces'.

In other cases the genitive structure is best transferred by an English sentence structure.

PERED NASTUPLENIEM KRIZISA
'before the occurrence of the crisis'

or, better,

'before the crisis occurs'.

The study of a large number of such cases suggests two patterns of investigation.

1. Some genitive government stretches are transformations of sentence structures, and some are not.

2. When certain relationships to sentence structures are observable in genitive nouns which might belong either in a noun-noun government stretch or in some other government stretch, they point to a definite order of precedence.

Analysis in Terms of Transformations

For analysis in terms of transformations, simple noun-noun government stretches consisting of only two members (N1, N2) were considered, since it has been found in all cases that multiple noun-noun government stretches consisting of more than two nouns in sequence can be resolved into a series of simple noun-noun structures, and that each of these can then be treated separately.

Where possible, each simple genitive government structure is transformed into a sentence structure. These simple structures are then classified according to the type of structure into which they are transformable. There are five classes.

1. A sentence structure featuring a transitive verb (Vx) in the reflexive form; the reflexive form is here equivalent to a passive. The initial noun is deverbal; the sequent noun is in the 'objective genitive' of traditional grammar.

OBSUJDENIE	TEZISOV	TEZISY	OBSUJDAHTS4
N1	N2	N2	Vx(N1)
The discussion of the theses.		The theses are discussed.	

2. A sentence structure featuring a transitive verb (V) in the non-reflexive form, or an intransitive verb, whether with a reflexive suffix or not. The initial noun is deverbal; the sequent noun is in the 'subjective genitive' of traditional grammar.

POSTANOVLENIE	PLENUMA	PLENUM	POSTANOVIL
N1	N2	N2	V(N1)
The resolution of the plenum.		The plenum resolved.	

3. A sentence structure of an equational type. The initial noun is de-adjectival.

VOZMOJNOST6	REALIZAQII	REALIZAQI4	-	VOZMOJNI
N1	N2	N2	is	A(N1)

The possibility of the realization. The realization is possible

4. A non-sentence structure featuring a preposition.

PROGRAMMA	POD7EMA	PROGRAMMA	PO	POD7E
N1	N2	N1	P	N2

The program of development. The program for development

5. No transformation is possible.

V R4DE RA1ONOV	V R4DE RA1ONOV
Nl N2	Nl N2

In a number of regions. In a number of regions.

Where transformation to a sentence structure is possible, the sequent noun comes to occupy the subject position and the initial noun the predicate position. (This fact is important for the transfer of the genitive ending to be discussed later.) Once these classes have been established on the basis of transformations, it may be possible to show that the possibility that a certain type of noun (deverbal, de-adjectival, etc.) will be followed by a genitive, and even by a genitive of a certain type (animate noun, deverbal noun, etc.) is greater or less. These facts would then form the basis of a system for determining which selection to make in an ambiguity and for judging whether a seeming noun-noun structure is actual or fortuitous.

The Transfer Difficulty

The genitive noun government routine is based on the assumption that in succession of two or more nouns where the sequents are in the genitive case, each sequent can be adequately translated by transferring the genitive ending I the English preposition 'of'.

But there are a number of cases in which the use of the preposition 'of' unsuitable, or, if suitable, very clumsy.

Consider this example.

PUTEM SOKRA5ENI4 VREMENI OBRA5ENI4 KAPITALA
'By means of curtailing of the time of circulation of the capital.

This was translated by the human translator as 'by curtailing the circulation time of capital'. The study of each noun-noun structure as rendered by the human translator proves interesting.

1. The human translator translated SOKRA5ENI4 VREMENI as 'curtailing the time'. The preposition 'of' is not used. The Russian is transformable into the sentence structure VREM4 SOKRA5AETS4 where SOKRA5AT6 is a transitive verb which happens to be reflexive (and therefore passive) in the transformation. This circumstance suggests the possibility that deverbal nouns from transitive verbs are to be translated as English gerunds and that the 'objective' genitive is to be transferred without rearrangement and with its genitive ending transferred as zero.

2. The human translator translated OBRA5ENI4 KAPITALA as 'circulation of capital'. The preposition 'of' is used. The Russian is transformable into the sentence structure KAPITAL OBRA5AETS4 where OBRA5AT6S4 is an intransitive verb which is reflexive as so many intransitive Russian verbs are. These circumstances suggest the possibility that deverbal nouns from intransitive verbs are to be transferred by the present routine.

3. The human translator translated VREMENI OBRA5ENI4 as 'circulation time'. The preposition 'of' is not used. The order of the nouns is reversed. (On occasion the noun which comes to stand first in the English is replaced by the corresponding adjective.) The Russian is not transformable into a sentence structure. These circumstances suggest the possibility that non-derived nouns (that is, not deverbal, not de-adjectival, etc.) are to be transferred by rearranging the order, and by transferring the genitive ending as zero. It is to be noted that provision must also be made for rearranging adjectives as in

```
3TI  DVIGATELI  VNUTRENNEGO  SGORANI4
A1   N1           A2           N2
```

which is to become

```
'these internal combustion engines'.
A1      A2       N2          N1
```

Another case is found in

```
PERED  NASTUPLENIEM  KRIZISA .
'before the occurrence of the crisis'
```

The human translator translated this as 'before the crisis occurred'. The preposition 'of' is not used. The Russian is transformable into the sentence structure KRIZIS NASTUPIL, where NASTUPIT6 is an intransitive verb. The English is rendered not as a preposition with a dependent noun-noun structure, but as a conjunction with a dependent clause. The basis for this is undoubtedly that the English preposition 'before', which is used here to translate DO, is homographic with the English conjunction 'before' and the occurrence of a deverbal noun made the transfer from prepositional phrase to clause very easy. (Mere homography is not sufficient, of course; certain semantic similarities must also be present.)

Where there is a combination of genitive structures, as in the first example, the question immediately arises as to the order in which these genii structures are to be treated. Experience suggests that they are to be treated in the order in which they have been discussed.

1. The head is a deverbal noun and the dependent is an objective genitive

2. The head is a deverbal noun and the dependent is a subjective genitive

3. The head is not a deverbal noun.

4. The head is a deverbal noun and is governed by a preposition whose English gloss is homographic with a conjunction.

Expansion of Coding

In view of the possibilities of resolving the syntagmatic difficulty and of resolving the transfer difficulty by the methods outlined here, it is proposed that nouns entered in the dictionary be coded to show whether they are transformable as heads of noun-noun structures. Five classes will be needed to cover both methods.

a) deverbal nouns from transitive verbs.
b) deverbal nouns from intransitive verbs.
c) de-adjectival nouns.
d) nouns capable of weak government of a noun through a specified preposition.
e) nouns which fall into none of the above categories.

SYNTACTIC ANALYSIS

The syntactic analysis is the operation which investigates the syntagmatic units which have been established and identifies the syntactic elements (subject and predicate) of each basic syntactic structure (machine sentence). (See the paper on the Sentence Separator Operation.)

The Basic Syntactic Structure in Russian

The basic Russian syntactic unit is composed of two elements. One syntactic element is the subject (here symbolized by H, for headword); the subject is described in the GAT as an independent variable. The other syntagmatic element is the predicate (here symbolized by P); the predicate is described in the GAT as a dependent variable which varies according to the subject. Once the subject and predicate are identified, the process of transferring the items surrounding them becomes relatively mechanical, since the majority of these items are in direct syntagmatic relationship with the two syntactic elements. Thus, all Russian sentence types are reducible to a subject construct and a predicate construct. Each construct may either be simple or be expanded by the three features of syntagmatic relationship: agreement, government, and modification.

The subject-construct nucleus, which, for the sake of brevity, will be called the subject (H), is in the nominative case. In a limited number of cases, the subject (the logical subject) is in the genitive case. The subject may or may not be present in a basic syntactic structure. The subject may be a multiple subject as a result of correlation of the type of 'eyes, ears, and nose', or of apposition of the type of: 'John, the baker'.

The predicate-construct nucleus, or predicate (P), may be a finite verbal form, a short-form adjective, a noun, or an adverb. Nominal clauses which act as predicates are to be classed as nouns, and prepositional phrases as adverbs. The predicate may or may not be present in a basic syntactic structure. The predicate may be a multiple predicate as a result of correlation of the type of: 'came and went', 'on the table or under it'.

If the digit 0 represents the absence of H or P, while the digit 1 represents the presence of one H or one P, and the digit 2 represents the presence of more than one H or P, every basic syntactic structure is of one of the following types.

0:0 no subject, no predicate
0:1 no subject, one predicate
0:2 no subject, more than one predicate
1:0 one subject, no predicate
1:1 one subject, one predicate
1:2 one subject, more than one predicate
2:0 more than one subject, no predicate

2: 1 more than one subject, one predicate

2:2 more than one subject, more than one predicate

The first symbol always refers to the subject and the second to the predicate, regardless of the order of the subject and the predicate in the Russian text.

The percentages of occurrence of the basic syntactic structure in machine sentences is approximately as follows. These figures are based on a corpus of more than 2, 000 text sentences.

0:0	0.5 %
1:0	12.0 %
0: 1	4. 5 %
1:1	56.0%
2:1	5.5%
1:2	5.5 %
2:2	16.0%

SYNTACTIC ANALYSIS

Subject Recognition

The subject (H) of a basic syntactic unit (machine sentence) is an independent variable; it may be present or absent in any given machine sentence, and, if present, may be single or multiple.

The subject is almost always in the nominative case, and a nominative case is almost always the subject. This fact gives a clear cue to the identification of the subject.

An unambiguous nominative is not always a definitive cue, however. A noun in the nominative case may be only one component of a multiple subject, and then the other components must also be found; a noun in the nominative case may also be the predicate, and, in that event, is no part of the subject at all.

Moreover, an unambiguous nominative is rare; most nominative forms are ambiguous in some way; the ambiguity most frequently involves the nominative, the accusative and the genitive cases. This adds considerably to the difficulty of establishing the subject, since any accusative not governed by a preposition is almost certain to be a part of the predicate, while any genitive not governed by a preposition may be part of an expansion either of the subject or of the predicate, or may even be the subject itself in certain cases. For all of the above reasons, the problem of establishing the subject is rather formidable.

The subject recognition operation is, however, perhaps the most important single operation in the GAT system. Continual efforts are being made to improve it and there have already been many modifications. Originally the subject identification operation was rather simple, and was based on the following assumptions.

Original Subject Identification Operation

There are three possible candidates for the office of subject:

(1) A noun in the nominative case;
(2) A noun in the genitive case;
(3) An adjective in the nominative case which is not in an agreement stretch with a noun.

(1) If i is a noun in the nominative case, and is either in the same sentence with a verb which agrees with it, or in a sentence without a verb, then i is a subject, and receives the code H in HPLOC. (The position known as HPLOC receives computer-generated codes which indicate the subject (H) and the predicate (P) of the machine sentence; this position

does not correspond to any particular position on the dictionary card The codes appear, however, in the printouts which give the details of the computer's operations.)

(2) If i is an adjective in the nominative case of the long form, and is not a member of an agreement stretch, then i is a subject, and receives the code H in HPLOC.

(3) If i is a noun in the genitive case and is not a dependent in a government stretch, but is in a sentence with NET, NE BYLO or NE BUDE then i is a subject, and receives the code H in HPLOC.

This routine was found to be only about sixty percent accurate, and so was expanded.

Expanded Subject Identification Routine

The subject routine tests every nominal form in the text. If the nominal is potentially nominative, it becomes a subject candidate.

If the subject candidate is unambiguously nominative, it is accepted as the subject of the sentence and the computer generates the code H in the HPLOC. If the nominative is actually the predicate and not the subject, the sentence will be recognized as having a multiple subject and no predicate (type 2:0). This situation will be resolved by suboperation 2:0 in the course of the syntactic analysis.

If the subject candidate is ambiguously nominative, it is most frequently potentially accusative as well as potentially nominative; (morphologically, a noun which is ambiguously nominative may also be potentially genitive, dative, locative, or some combination of these, certain adjectives which are ambiguously nominative may also be potentially instrumental; but these ambiguities will usually have been resolved by the case cleaning routines.) A series of three basic tests is used to determine which of the subject recognition suboperations is to be used.

1. Is i a member of a government stretch?

2. Is i-1 a form which determines the accusative case?

3. Is i-1 a member of a prepositional government stretch?

These three questions lead into certain subroutines, depending on the answers given. These subroutines may be summarized as follows.

		Y	N	A
A				
1.	Is i coded as a member of a government stretch?	B	D	-

		Y	N	A
B				
1.	Is the stretch a noun-government stretch (1122)?	2	C	-
2.	Is i the initial of the stretch?	3	4	–
3.	Then i is assumed to be ambiguously nominative or accusative only, since any other potential case will either have been established to the exclusion of all others or have been discarded during the case cleaning operation. Tests are then made to establish whether there is an item which governs that noun in the accusative. If there is, i is not the subject; if there is not, it is coded as the subject.	-	-	Z
4.	Is i the final of the stretch?	5	6	–
5.	Then i is assumed to be ambiguously nominative, accusative and genitive, since, if it were not genitive, it could not have been recognized as a member of the 1122 stretch. If there is at least one medial noun in the stretch as it stands, tests are made to establish whether there is a plural verb of which i can be the subject, and whether there is also another noun, similarly ambiguous, which can also be a subject candidate. If there is a plural verb, but no other potential nominative plural, i is coded as the subject.	-	-	Z
6.	It is assumed that any medial noun in a 1122 stretch is in dubitably in the genitive case, and i is therefore not a subject.	-	-	Z

		Y	N	A
C				
1.	Is i a member of a government stretch other than a noun-government stretch?	2	D	-
2.	It is assumed that such a noun cannot be a subject and i is therefore not a subject.	-	-	Z

D

1.	Is i+1 an item which governs the accusative?	2	E
2.	It is assumed that the subject candidate is ambiguously nominative, accusative, or genitive. If nominative, it is a subject; if genitive, it may be the logical subject, since it is not in a government stretch; but if it is accusative, it cannot be a subject. Tests are therefore made to determine whether i is in the accusative.	-	-
3.	Is i-1 an item which governs the accusative?	4	5
4.	It is assumed that if both i-1 and i+1 govern the accusative, i must be accusative, since a noun would not occur in such a position unless it were governed by i-1, and so i is not a subject.	-	-
5.	It is assumed that, if i-1 governs the accusative, but i+1 does not, i may be a subject, and certain tests are carried out which involve establishing the identity of the items i-1 and i+1. These tests may or may not lead to the conclusion that i is the subject.	-	-

E		Y	N	A
1.	Is i-1 a member of a prepositional government stretch (and therefore the final, since it is already established that i is not a member of any government stretch)?	2	3	-
2.	It is assumed that i is a subject.	-	-	Z
3.	A battery of similar tests, involving the function of i-1, i-2, i-3, i+3 and i+4 is carried out, and, on the basis of criteria very similar to those already described, a decision is made as to whether i is a subject. Once it has been decided that i is a subject, the final test is to compare it with the potential predicate to try whether the subject and predicate are in agreement.	-	-	Z

Z		Y	N	A
1.	Are there any remaining subject candidates?	2	99	
2.	Call the next subject candidate i.			-A

The operation described above is approximately seventy-five percent effective. Since such a percentage is still too low, continuous attempts have been made to improve the subject recognition operation. The effectiveness has now been raised above eighty percent.

Recent Trends in Subject Recognition Research

The latest improvements focus on the probable position of the subject in the sentence. A subject candidate which is the initial item of the sentence is accepted as a subject. Otherwise, searches are made on the assumption that the subject is most probably the second or third item of the sentence; for this purpose, all well-defined syntagmatic structures, such as prepositional phrases, are considered to be one item. There is a growing feeling that it will be possible to recognize the subject definitively by this method.

On the other hand, the suggestion has been made that a search for the predicate, if there is a clearly predicative form in the sentence, would considerably shorten the search for the subject. Information about the predicate would indicate whether the subject is to be singular or plural (plural includes a multiple subject consisting of a number of singulars or plurals); ambiguities between genitive singular and nominative plural, for example, need not be investigated if the predicate indicates that a nominative singular subject is to be expected.

Another suggestion has been made to the effect that, while Russian word order is free, it is not completely haphazard, and that some general principles can be evolved as to whether the subject is more probably to be found to the left or to the right of certain types of verb.

It has been argued, however, that a great number of the cases where the subject is not adequately recognized at present are cases of ambiguity in the part of speech (is CTO a conjunction or a pronoun, and if a pronoun, is it nominative or accusative?) or of the absence of any subject, and that the above suggestions would not materially improve the output. Tests to determine the relative value of these arguments are now being prepared.

Predicate Recognition

There are three possible candidates for the office of predicate:

(1) a finite verb
(2) a short-form adjective or participle
(3) an adverb.

(It has been indicated above that a noun in the nominative may also be predicate. For the time being, such a predicate noun is coded as a subject)

(1) If i is a finite verb, it is a predicate, and receives the code P in HPLOC

(2) If i is a short form adjective or a short form participle and i-1 is not a finite verb, then i is a predicate and receives the code P in HPLOC.

(3) If i is an adverb which has the code P in the nominative singular location (III:66), then i is a predicate and receives the code P in HPLOC. (These adverbs are eighteen in number; they almost always occur as predicates when they occur at all: NADO, NEL6Z4, VOZMOJNO, etc.)

Certain other tests for predicates are used. These tests were devised very early and on the basis of only the chemical text. Their usefulness in a wider field is open to doubt, and research is now being conducted to determine whether it is necessary to modify them.

On the whole, the recognition of predicates is not nearly as involved as the recognition of subjects, and the operation is almost completely effective except in the recognition of short-form adjectives which have the same form as adverbs.

SYNTACTIC ANALYSIS

Sentence Type Operations

Each syntactic unit (machine sentence) is inspected for subjects and predicates, and a count of subjects and of the predicates is made.

On the basis of this count, the sentence is assigned to one of the nine sentence types. For each sentence type there is a suboperation which analyzes it. This suboperation is designated by the same combination of numbers as the sentence type.

Thus if one subject and one predicate are found in a sentence, the sentence is of type 1:1, and its analysis is effected by suboperation 1:1. Since there are nine sentence types, there are nine suboperations, although some of them are brief almost to the point of non-existence.

The suboperations for the various syntactic types have two chief purposes:

a) Syntactic units which are ostensibly of one syntactic type are analyzed to test whether they are not really a combination of other syntactic types. If they prove to be so, the units are then modified, or sentence separator boundaries are introduced to separate the components into different syntactic units.

b) The syntactic units which are definitely established are analyzed to determine the point of division between the subject and the predicate (dichotomy cut). This point of division is indicated by the storing of the machine-generated code D in DCLOC of the predicate item nearest to the subject, which is the first item of the predicate if the subject precedes, and the last item of the predicate if the subject follows. (The position known as DCLOC (dichotomy cut location) does not correspond to any particular position in the dictionary cards. The code D appears,however, in the printouts which give the details of the computer operations. This code D is also said to be stored in DINDA (definite and indefinite article location) in some descriptions.)

These two chief purposes of the suboperations are demonstrated in the descriptions which follow.

In this description, there are a number of instructions to insert specific words before a predicate. This is effected by establishing the lefthand boundary of the predicate construct after including all dependent modification or agreement stretches, and by inserting the specific words in the EPREP and DINDA areas of the leftmost item, and to the left of any other words inserted or to be inserted there. (EPREP (English preposition) and DINDA (definite and indefinite articles) are two contiguous areas of the work space and may be thought of as corresponding to positions II:36-34 on the dictionary cards, that is, to the unused Positions immediately to the left of the first English gloss. The area is not limited to exactly that number of positions, however, since many of the positions to the right of II:44 are usually vacant, and therefore usable, while, in any

case, the area may be made larger by repositioning the material on either side of it.)

	A. Counting	Y	N	A
1.	Is there a sentence to the right?	2	99	
2.	Determine the number of subjects in the sentence and register it.	-	-	
3.	Determine the number of predicates in the sentence and register it.			
4.	Combine the numbers in the order in which they were determined.	-	-	
5.	Move to the suboperation designated by that combination of numbers.	-	-	

	B Suboperation 0:0			
1.	No operation is necessary.	-	-	Z

	C. Suboperation 0:1			
1.	Does the particle NE occur in the syntactic unit?	2	4	—
2.	Is there a nominal in the genitive case?	3	4	—
3.	Mark this genitive nominal as the subject.	-	-	F
4.	Is the predicate a finite verb?	6	14	
5.	Is the predicate in the first person plural?	6	7	-
6.	Insert 'let us' before the predicate construct.	-	-	Z
7.	Is the predicate transitive?	8	11	-
8.	Is there a nominal in the accusative case?	9	*	-
9.	Mark this accusative nominal as the subject.	-	-	10
10.	Mark the verb for transfer as a passive verb in English.	-	-	Z
11.	Is the predicate singular?	12	13	- -

		Y	N	A
,2.	Insert 'it' before the predicate construct.			Z
13t	Insert 'they' before the predicate construct.	-		Z
14.	Is the predicate a short-form participle or a short-form adjective?	15	22	
15.	Is i-1 KAK?	16		19
16.	Is the predicate singular?	17		18
17.	Insert ' is ' before the predicate construct.			Z
18.	Insert ' are ' before the predicate construct.			Z
19.	Is the predicate singular?	20		21
20.	Insert 'it is' before the predicate construct.	-	-	Z
21.	Insert 'they are' before the predicate construct.	-	-	Z
22.	Is the predicate an adverb?	23		*
23.	Mark the adverb for transfer as an adjective.	-	-	24
24.	Insert 'it is' before the predicate construct.	-		Z

— Suboperation 0:2

1.	Is there a comma or a conjunction between any two predicates?	2	Z
2.	Mark a sentence separator boundary at that comma or conjunction.	-	Y

E Suboperation 1:0

1.	No operation is necessary.	-	Z

— Suboperation 1:1

1.	Is the subject a numeral?	2	5-
2.	Is the numeral 1?	3	4 –

3.	Mark the predicate as singular.	-	-	
4.	Mark the predicate as plural.	-	-	
5.	Is the subject to the left of the predicate?	6	8	
6.	Establish the predicate construct boundary i-n.	-	-	
7.	Store the dichotomy-cut code D in DCLOC of i-n.	-	-	
8.	Establish the predicate construct boundary i+n.	-	-	
9.	Store the dichotomy-cut code D in DCLOC of i+n.	-	-	10
10.	Is the predicate a finite verb?	11	13	
11.	Is the predicate reflexive?	12	Z	
12.	Insert 'there' before the predicate construct if the boundary is 1+n, otherwise not.	-	-	Z
13.	Is the predicate a past passive participle?	14	17	
14.	Is the participle singular?	15	16	
15.	Insert 'was' before the predicate construct.	-	-	Z
16.	Insert 'were' before the predicate construct.	-	-	I Zj
17.	Is the predicate an adjective?	18	21	-
18.	Is the adjective singular?	19	20	
19.	Insert 'is' before the predicate construct.	-	-	Z
20.	Insert 'are' before the predicate construct.	-	-	Z
21.	Is the predicate an adverb?	22	26	
22.	Is the subject singular?	23	24	
23.	Insert 'is' before the predicate construct.	-	-	25
24.	Insert 'are' before the predicate construct.	-	-	25
25.	Mark the adverb for transfer as an adjective.	-	-	Z
26.	Is the predicate NET?	27	33	

		Y	N	A
27.	Is the subject a genitive noun?	28	0	-
28.	Mark NET for transfer as zero.	-	-	29
29.	Is the subject singular?	30	32	
30.	Insert 'there is no' before the subject construct.	-		31
31.	Mark the subject as accepting no article insertion (code 9 in I:78).	-	-	Z
32.	Insert 'there are no' before the subject construct.	-	-	Z
33.	Is the subject a noun in the genitive?		34	38
34.	Does NE occur?		35	38
35.	Insert 'no' before the subject construct.	-	-	36
36.	Mark NE for transfer as zero.	-	-	37
37.	Mark the subject as accepting no article insertion (code 9 in I:78).	-	-	Z
38.	Is the subject CTO?	39	0	-
39.	Mark the predicate as singular.			Z

G. Suboperation 1:2

1.	Do all predicates follow the subject?	2	7	-
2.	Are the predicates of the same number and person as the subject?		3	4 -
3.	Store the dichotomy-cut code D in HPLOC of the left most item of the first predicate.	-	-	Z
4.	Is the first predicate of the same number and person as the subject, but some subsequent predicate not?		5	0
5	Is there a comma between the first and second predicate?		6	Z
6	Mark this comma as a sentence separator boundary.	-		Y
7	Do all predicates precede the subject?	8		1 3

		Y	N	-
8.	Are the predicates of the same number and person as the subject?	9	10	
9.	Store the dichotomy-cut code D in HPLOC of the right most item of the last predicate.		Z	
10.	Is the last predicate of the same number and person as the subject, but some prior predicate not?	11	0	
11.	Is there a comma between the predicate which is in agreement with the subject and the predicate which is not ?	12	Z	
12.	Mark this comma as a sentence separation boundary.	-	-	Y
13.	Do the predicates occur both to the left and to the right of the subject?	14		
14.	Call the first item of the machine sentence i.	-	15	
15.	Is there, to the right of i, a comma, a conjunction, or an adverb which has the code P in III:66, between predicate and predicate or between predicate and subject?	16	Z	
16.	Call this comma, conjunction or adverb i.	-	-	1?
17.	Is i+1 a subject?	27	18	
18.	Is i+1 a predicate?	27	19	
19.	Is i+1 a member of an agreement stretch (31IX)?	20	22	
20.	Is i+2 a subject?	27	22	
21.	Is i+2 a predicate?	27	22	
22.	Is there an item which is BOTH a member of an agreement stretch and is a subject?	27	23	
23.	Is i+1 a member of a modification stretch (433P)?	24	25	
24.	Is i+2 a subject?	27	25	
25.	Is i+1 a member of a modification stretch (423P)?	26	15	
26.	Is i+2 a predicate?	27	15	
27.	Mark i as a sentence separator boundary.	-	Y	

	Y	N	A
H- Suboperation 2:0			
1. Is there a comma or a conjunction between any pair of subjects?	2	3	
2. Mark that comma or that conjunction as a sentence separator boundary.	-		Y
3. Is the first subject singular?	4	5	
4. Insert 'is' before the second subject construct.			Z
5. Insert 'are' before the second subject construct.	-	-	Z
I. Subroutine 2:1			
1. Is the next to the last word in the machine sentence an adjective or an adverb?	2	5	-
2. Mark this adjective or adverb as a predicate.			J
3. Does the predicate precede all subjects?	4	5	-.
4. Insert 'there' before the predicate-construct.	4	5	
4. Is there a comma, conjunction or dash between subject and subject?	6	Z	
6. Store a dichotomy-cut code D in DCLOC of the left most item of the predicate construct.	-	-	Y
J_. Suboperation 2:2			
1. Is there a subject (i-n) and a predicate (i) and a comma or conjunction (i+n)?	2	Z	.
2. Mark the comma or conjunction as a sentence separator boundary.	-	-	Y
Y. Movement			
1. Move left to the nearest sentence separator boundary.	-	-	A

Y__N

Z. Movement

1. Move to the sentence separator boundary to the right. - 1

LEXICAL CHOICE

Many Russian words require more than one English gloss; it is possible to have a one-to-one correspondence between entry and gloss only in the case of highly technical terminology. If an even passable translation is to be achieved, there must be multiple glosses and a method of choosing among them. The GAT method of choosing among the glosses is called Lexical Choice.

The GAT dictionary contains space for two English glosses. Any other glosses must be stored in ancillary dictionaries. Such storage requires greater memory space; it also requires multiple searchings which increase the computer time. In addition, it is difficult to evolve lexical choice operations of any wide generality if ancillary dictionaries are used. For all of these reasons it is desirable, from the technical point of view, that there be no more than two glosses to an entry. Only a relatively small proportion of the dictionary entries have more than two glosses at present. Some prepositions, however, have as many as sixteen.

Lexical Choice Codes

The codes that are used in the lexical choice operations are of four types. There are a) the lexical choice candidacy code, b) the lexical choice review code, c) the lexical choice operation codes, and d) the environment determining, or identification, codes.

a) The lexical choice candidacy code is the code 1 in position I:74; it indicates that the entry is a candidate for lexical choice.

b) The lexical choice review code is the code 4 in position I:75; it indicates that the entry probably requires lexical choice coding. The review code is used whenever the press of time or other work does not allow the fairly complex lexical choice analysis to be made immediately. When time becomes available, the entries which require lexical choice analysis can be found quickly by searching for this code, and both the analysis and the coding can be completed.

c) The lexical choice operation code is a three-position code located in I:76-78. Each code designates the particular operation to be used in determining the choice of the English gloss. Almost all such operations are applicable only to a particular entry, but a few are more general. Both general and particu lar operations are exemplified later.

d) The choice in a lexical choice operation frequently depends on the character istics or identity of textwords in the environment. If, in any text, a word has been found to determine an environment (and thus to indicate a lexical choice in some other word), it receives an environment determining code.

The environment determining codes are best discussed under four heads either because of differences in position in the dictionary (i and ii below), or because codes originally designed for other purposes were found to also useful as environment determining codes. (iii and iv below).

(i) The basic environment determining code is a two-position code located in I:47-48. (The exclusion codes also occur in these positions. But no item which is excluded is also a candidate for lexical choice, and so there is no overlapping of codes.) It is some-times useful to indicate whether the environment determining code in I:47-48 accompanies a split or an unsplit dictionary entry. This is shown in writing by enclosing it in parentheses if it is found only in the split dictionary, by enclosing it in square brackets if it is found only in the unsplit dictionary, and by enclosing it in both square brackets and parentheses if it is found in both dictionaries.

(ii) Other environment determining codes are located in I:74-75. This location of coding in these positions is only possible when the environment determining item is not also a lexical choice candidate.

(iii) Since most lexical choice operations are associated with one particular word or family of words, the lexical choice code numbers of these operations can also serve as identification codes for those words or families of words. Consequently, if some other lexical choice operation is being applied, and if it becomes necessary to identify an item . which is a lexical choice candidate as being in the environment, the lexical choice operation number of that item can serve as an environment determining code.

(iv) Idiom codes also uniquely identify the words that carry them and so can serve as environment determining codes when necessary.

The Lexical Choice Procedure

If an item is a candidate for lexical choice, the lexical choice operation number indicates which operation is to be used. This operation may check on the presence of certain words in the environment because these words determine an environment in which a given lexical choice is to be made. Such words can be identified either by their environment determining code or by their specific shape as a textword, although the latter method is relatively complicated. The lexical choice operation then directs the computer to the first gloss, to the second gloss, or to one of the ancillary dictionaries for other glosses. The gloss thus selected is the one used in the translation.

Sometimes it is best not to translate an item at all. Thus, the two Russian words SBOR UROJA4 are most conveniently translated by the one word 'harvest'. This means choosing the gloss 'harvest' for SBOR and choosing no gloss for UROJA4. In such a case, the idiom candidate code position (I:49)

receives the computer-generated code X. This code will prevent the synthesis of UROJA4, and so will avoid a situation where the textword is not translated, but its genitive singular ending is transferred by a corresponding English structure (in this case, probably, by the preposition 'of').

Lexical choice operations vary so much that they cannot be described in any general way. Examples of individual operations make the best description.

Most lexical choice operations apply only to one word. There are a few which can apply to more than one word.

The following operations are applicable to more than one word.

Operation 229	Y	N	A
1. Does i have the reflexive suffix (code 3 in IV: 37)?	2	3	
2. Choose gloss 2.	-	-	99
3. Choose gloss 1.	-	-	99

Examples:	Russian Word	Gloss 1	Gloss 2
	PROIZVODIT6	'effect'	'be carried out'

Operation 303	Y	N	A
1. Is i a predicate (machine-generated code P in HPLOC) ?	2	3	–
2 Choose gloss 1.	-	-	99
3. Choose gloss 2.	-		99

Examples:	Russian Word	Gloss 1	Gloss 2
	VIDNO	'evident'	'evidently'

Operation 208	Y	N	A
1. Is i+1 in the instrumental case?	2	3	.-
2. Choose gloss 1.	-	-	99
3. Choose gloss 2.	-	-	99

Examples:	Russian Word	Gloss 1	Gloss 2
	OTLICAT6S4	'be characterized'	'differ'
	PREDSTAVIT6	'represent'	'present'
	SLUJIT6	'serve	'word'

The following operations apply to only one word. In these cases the Russian example is given before the statement.

Russian Word	Gloss 1	Gloss 2
VODY	'water'	no translation

Operation 102

		Y	N	A
1.	Is i-1 PAR- (recognized by environment determining code BC in I:47-48)?	2	3	
2.	Choose gloss 2 and place code X in I:49.	-		99
3.	Choose gloss 1.			99

Russian Word	Gloss 1	Gloss 2
PAR	'vapors'	'steam'

Operation 111

		Y	N	A
1.	Is i plural?	3	2	–
2.	Choose gloss 1.	_	_	99
3.	Is i+1 VODY (recognized by its own lexical choice operation code 102)?	4	5	–
4.	Choose gloss 1.	_	_	99
5.	Choose gloss 2.	_	_	99

Russian Word	Gloss 1	Gloss 2
SLEDUET	'be necessary'	'follow'

Operation 227

		Y	N	A
1.	Are i+1 and i+2 respectively a comma and CTO (lexical choice number 607)?	2	3	–
2.	Choose gloss 2.	_	_	99
3.	Choose gloss 1.	_	_	99

Russian Word	Gloss 1	Gloss 2	Gloss 3
MEN6WE	'smaller'	'less'	'slower'

Operation 409

		Y	N	A
1.	Is i+1 SKOROST- (CodeCN in I:47-48)?	2	3	–
2.	Choose gloss 3.	.	_	99
3.	Transfer to Operation 303.	-		99

Russian Word	Gloss 1	Gloss 2
A	'but'	'and'

Operation 601	Y	N	A
1. Is i-n NE (lexical choice 704)?	3	2	—
2. Choose gloss 2	-	-	99
3. Is there a mark of punctuation between i and i-n?	4	5	—
4. Choose gloss 1.	-	-	99
5. Choose gloss 2.	-	-	99

Some lexical choice routines are longer than the samples given, and provide for as many as sixteen glosses. But they are rarely more complex.

Routine 704

The routine for the negative particle NE is one of the most complicated lexical choice routines. It is given here both as an example of complexity and because it is helpful in understanding the description of English Synthesis which follows.

It should be noted that the synthesis operations for participles as well as the lexical choice operation for the particle BY include a treatment of the particle NE. Thus, when either BY or a participle occurs in the environment of NE, the lexical choice operation for NE is not used.

A.	Y	N	A
1. Is i+1 a verb?	2		
2 1 –			
2. Is i+1 an infinitive?	21	3	
3. Does the gloss of i+1 contain the verb 'be' (code X in I:73)	B	4	
4. Is i+1 a participle?	99	5	
5. Is i+1 present tense?	6	15	-
6. Is i+1 a reflexive verb (code 3 in I:68)?	7	12	-
7. Is i+1 to be transferred as an English passive (Code 1 in I:40)?	8	1	3 -
8. English Past Participle Synthesis: Synthesize the participle.			9
9. Is i+1 singular?	10	24	
10. Is i+1 first person?	22	11	
11. Is i+1 third person?	23	24	
12. Is i+l any form of the verb MOC6 ?	30	13	
13.. Is i+1 singular?	14	26	
14.. Is i+1 third person?	25	26	

		Y	N	
15.	Is i+2 BY?	99	16	
16.	Is i+1 a reflexive verb (code 3 in I:68)?	17	29	
17.	Is i+1 to be transferred as an English passive (Code 1 in I:40)?	18	29	
18.	English Past Participle Synthesis: Synthesize the participle.			
19.	Is i+1 singular?	20	28	
20.	Is i+1 second person?	28	27	
21.	Transfer i as 'not'.	-	-	99
22.	Transfer i as 'am not'.	-	-	99
23.	Transfer i as 'is not'.	-	-	99
24.	Transfer i as 'are not'.	-	-	99
25.	Transfer i as 'does not'.	-	-	99
26.	Transfer i as 'do not'.	-	-	99
27.	Transfer i as 'was not '.	-	-	99
28.	Transfer i as 'were not'.	-	-	99
29.	Transfer i as 'did not'.	-	-	99
30.	Transfer i and i+1 together as 'cannot'.	-		99

B.

		Y	N	A
1.	Is i+1 BYLO?	2	6	–
2.	Is there a nominal structure (n) in the genitive case and not in a government stretch (X122)?	3	16	-
3.	Is n in an agreement stretch (3112) ?	4	5	–
4.	Call the first item of this agreement stretch n.	-	-	5
5.	Is n singular?	12	13	-
6.	Is i+1 in the past tense?	7	9	–
7.	Is i+1 singular?	8	17	-
8.	Is i+1 in the second person?	17	16	-
9.	Is i+1 singular?	10	18	-
10.	Is i+1 second person?	18	11	-
11.	Is i+1 third person?	19	20	-
12.	Transfer i as 'there was no'.	-	-	14
13.	Transfer i as "there were no'.	-	-	14
14.	Transfer i+1 as ZERO.	_	-	15
15.	Mark n as not accepting any article.	-	-	99
16.	Transfer i and i+1 as 'was not'.	-	-	99
17.	Transfer i and i+1 as 'were not'.	-	-	99
18.	Transfer i and i+1 as 'are not'.	-	-	99
19.	Transfer i and i+1 as 'is not'.	-	-	99
20.	Transfer i and i+1 as 'am not'.	-	-	99

RESEARCH SEMINARS ON SEMOLOGY

Research seminars on semology were conducted by the Georgetown University Machine Translation Project between 1960 and 1963. A discussion of these seminars may be set under three heads: the research problem and methodology, a summary of the discussions, and the conclusions.

The Research Problem and Methodology

Efficient and effective ways of handling polysemy, "idioms", a phrase meanings, contextual effects, and connotative meanings--to name but a few aspects of the problems of translation-- has always been a basic need in machine translation. The Georgetown Machine Translation Research Project approaches these problems linguistically, with the aim of discovering structural cues to the meaning in the source and target language themselves. The discovery of such structures will, it is felt, lead to the formulation of efficient rules for programming. The first efforts in machine translation were necessarily concentrated on the compilation of the basic dictionaries, and on the development of analyses at several structural levels. When these had been largely developed, however, it was seen that the translation results, usable and even very good though they were, required further refinement. In most cases, such further refinement depended on the development of ways of treating semantic problems.

It was decided to hold a series of research seminars in what has since come to be called semology. At the seminar sessions the Georgetown staff and the consultants, (Dr. Martin Joos and Dr. George L. Trager) discussed material involving various semological problems, formulated tests and experiments, examined the results, and came to conclusions which have contributed, or will eventually contribute, to the solution of the problems studied.

This series of seminars has shown that research into semological structures is possible, and that it produces both important theories on procedure and practical improvements in translation.

Summary of Discussions

Seminar sessions were held about once a month, summers excluded, from October 1960 until June 1962. After that meeting Dr. Trager, who had established himself temporarily in New Mexico, was unable to attend regularly. During the summer of 1962 a committee of the staff, without consultants, met daily from June 27 to August 16. Three seminar sessions were held with Dr. Joos in the fall of 1962, but ill-health made it impossible for Dr. Joos to continue. A meeting took place in January 1963; Dr. Trager attended and drew up the report which serves as the basis for this paper.

Detailed reports on the work done and the conclusions reached exist in the form of the recorders' notes on each session.

The first session was begun with a report by Professor Dostert on the status of the Georgetown Machine Translation Research Project at that time. A Russian text in the field of organic chemistry had been translated, and the translation had been evaluated by a chemist as adequate and useful. A keypunch center had been established in Germany. The 705 program was being converted to the 709 program. Syntactic research has increasingly pointed up the need for semantic research.

A report on data indexing procedure as developed at NSA followed, with discussions of the relevance of these procedures to machine translation. It was concluded that there are relationships which merit attention in future research.

The terminology being used in the Georgetown Machine Translation Project was considered, and suggestions for refining definitions and for improving applications were made.

Subjects for the seminar series were suggested, and the series was definitely oriented toward semological research at this session.

In preparation for the November seminar, a Russian text was distributed with questions about the insertion of articles in the English translation, about phrase structure, and about the limitations of prepositional phrases. The study of the insertion of articles proved to be of great theoretical importance, and much attention was devoted to it from this time on.

At the seminar in November, patterns for the insertion of the articles in began to emerge, and these were clearly beginning to be characterized as meta-syntactic, that is, as semological. The discussion carried over to the December meeting, with additional preparatory study of various kinds of material. At that session the further suggestion came out that some uses of the articles in English are even beyond semology; they are metalinguistic.

At the February and March meetings discussion of the article continued, and various concrete decisions were made for testing the tentative conclusions. ' A point that was made was that the structure and vocabulary of the target language must be statistically adequately mastered by the machine for adequate and acceptable output, whereas the knowledge of the source language can be less technically precise. This is an analogy to human translation which has not always been recognized.

The April meeting was devoted to discussion of several other research problems — the use of English passives, the concordance lists, and the possi bility of systematizing the use of hyphens as a guide to certain kinds of structures. A report on recent achievements in solving specific problems was given.

At the May meeting a further such report, in great detail, was received very favorably by the seminar group. It dealt with a basis for semological coding by recognition of semological congruences. This was concrete evidence of progress in analysis and application resulting directly from the seminar series.

The June meeting dealt with plans for the September International Conference on Machine Translation, and also included several technical reports and a summary of progress during the year.

After the summer break, seminar sessions were resumed in September 1961. There was first a report on the International Conference, and sets of agenda were projected for subsequent meetings.

The plan was to deal with general semological questions, to make more precise the structural dividing line between syntax and semology, and to present technical reports on aspects of Chinese structure and on the Slavic languages as a group.

At the next few meetings the Comparative Slavic Research was discussed in various ways. A preliminary report was given, and some technical results were presented. It appeared most efficient to use Russian as the base language in this particular field.

In January 1962, there was a presentation by Dr. Joos of some problems in the statistics of vocabulary.

At the February and March sessions, the English articles were again discussed and concrete conclusions were reached on the basis of an experiment implemented by Dr. Trager. The experiment involved having native speakers of English reinsert articles that had been removed from a text with whose area of discourse the speakers were certain to be familiar. It appears that while about ninety percent of the articles in a text can be inserted by a suitable syn tactic analysis, and about ninety percent of the remaining articles can be inserted by semological analysis, there remains a residue of one or two per cent that is beyond machine translation. This residue, consisting of the uses that are metalinguistic, that reside "in the mind of the user", is not subject to prediction or to systematization.

In May, Dr. John de Francis presented a report on preliminary work directed at the machine translation of Chinese.

At the June seminar sessions there was a technical report on certain Russian structures successfully analyzed for translation into English. This was followed by an extended discussion of guidelines for future semological research.

During the summer of 1962, sessions were held almost daily, a great many technical reports were presented and discussed, and useful conclusions Were drawn. A general and recurrent conclusion was that many apparently

semological problems are really within the area of syntactic structure if the various structural levels are sufficiently exploited. Further research which would distinguish problems as to the linguistic level on which they exist was proposed.

In October 1962 there were two sessions. At the first, the work done during the summer was reviewed, and some concrete guidelines were laid down for further investigation. At the second, the details of a procedure for thesaurus coding were discussed.

In November, reports were given on various activities in Europe and technical presentations were made.

The session in January 1963 was devoted to a review of all the work and to the planning of this summary.

Conclusions

The Machine Translation Research Project seminars developed a great deal of useful and precise information, both theoretical and practical, which will tend to make machine translation more effective and efficient.

It is definite that level-by-level analyses of the structures of the source and target languages must be made. Only by such structural analyses can the practical rules of operations be simplified and condensed into efficient programs.

It has been demonstrated that careful vocabulary listing, and intensive grammatical coding (including syntactic coding) will take care of a large numb* of seemingly semological problems. When this work is complete, it will clearly show the limits of its effectiveness, and so pave the way for more effective semological analysis.

It has been shown that semological structure can be described and analyzed, and several promising procedures for such analysis have been proposed.

Finally, it has been possible to see that a point is reached where meta-linguistic considerations must enter in.

Prospect:

Future research and activity in the field of MT would seem to be most useful along the following lines:

1) Further systematization, with a view to simplification in application of vocabulary listing procedures and of grammatical (morphemic and syntactic) analyses;

2) Extensive development of semological coding within the (micro-) linguistic systems;

3) An approach to metalinguistic analysis by studies of the cultural system in general, including analyses of style and content along "language-and-culture" lines.

Further Research

A more generalized system for making lexical choices is the object of the immediate research at Georgetown. Researchers have already mapped different paths of attack and are exploring them. It is hoped that lexical choice can be made to depend on the purely formal characteristics of the text, or, failing that, on a limited system of broad semantic coding.

Analysis in Terms of Semantic Order Categories

On such path of attack has been explored by M. Zarechnak.

A preliminary analysis of approximately 10, 000 noun-government structures in the genitive indicates that nouns can be classified in terms of their order in complex genitive structures.

T

These classes are here described as semantic; once the possibilities of co-occurrence and relative order have been established, the classes seem most easily definable on a semantic basis, since there are certain semantic components common to the definition of all members of each class. Thus each class is here described by a rubric which epitomizes the semantic characteristics of the class.

Thus the members of one order class can be subsumed under such a rubric as "quantifier". A quantifier is any noun which refers to quantify, a whether vague or precise, whether in terms of individuals or in terms of mail

The members of two other order classes may be subsumed under the rubric "partitive". A partitive is any noun which refers to a part of a whole. Partitives may be structured or unstructured. A structured partitive refers to a part of the whole which is fixed in form or volume, such as half of a pie, or the leg of a table. An unstructured partitive refers to a part of the whole which has no fixed form or volume, such as a piece of pie, or a chunk of wood.

Another rubric used to describe an order class is "qualifier". A qualifier is any noun which refers to a quality, such as beauty or value.

Still another rubric is "process noun". Process nouns are either transitive or intransitive. A process noun, intransitive, is a deverbal noun derived from an intransitive verb. A process noun, transitive, is a deverbal noun derived from a transitive verb.

The order structuration of semantic components can be illustrated by using the sub-class of inanimate concrete nouns like STOL: 'table' as the nuclear element. (Such a noun is inanimate because its accusative plural

has the same form as its nominative plural; it is concrete because it occurs in such frames as 4 VIJU_____ : 'I see the _____ '). Such a noun elicits certain predictable sequence patterns when in association with other nouns.

It is possible to describe in definite terms the order of nouns about a concrete inanimate noun. The position of the concrete inanimate noun is given as zero, and negative numbers represent the order of precedence to the left of that noun, and positive numbers the order of precedence to the right. These precedence numbers do not imply any fixed distance from the noun under consideration; they indicate the order which the appendant nouns assume with relation to each other, the nouns with smaller precedence numbers being always nearer to the nucleus than those with larger precedence numbers, however many or few nouns actually appear.

Position	Rubric for the Semantic class of the noun
-12	Negation of existence
-11	Affirmation of existence
-10	Space
- 9	Process, transitive
- 8	Process, intransitive
- 7	Any number
- 6	A structured quantifier referring to number (plural only)
- 5	A structured quantifier referring to group or mass (mass only)
- 4	(A parenthetical sentence structure)
- 3	A qualifier in displaced order
- 2	An unstructured partitive quantifier
- 1	A structured partitive quantifier
0	THE NOUN UNDER CONSIDERATION
1	Color
2	Name of color
3	Colored
4	A quantifier in displaced order
5	A qualifier

A quantifier is in displaced order when it occurs after the noun under consideration instead of in its expected position before it. A qualifier is in displaced order when it occurs before the noun under consideration instead of in its expected position after it.

Examples:

NITKA	JEMCUGA	QVETA	OXOTNIC6E1	KARTECI
- 1	0	1	(adjective)	2

A string of pearls of the color of shotgun pellets

GRUDA NERA ZOBRANNYX OBLOMKOV NAMERZWEGO L6DA
-5 (Adj.) -2 (Adj.) 0
A load of jumbled chunks of ice frozen together

DVA LOMT4 CERNOGO XLEBA
-7 -2 (Adj.) 0
Two loaves of black bread

FAKT POLUCENI4 ORDENA
-11 -9 0

The fact that the order was received

TYS4CA KUBOMETROV IVY, V4ZA, LIPY
- 7 - 6 0 0 0
A thousand cubic meters of willow, elm, and linden

TE JE DVA, KAZALOS6, LOMT4 PRORJAVEVWE1 SELEDKI
(Adj.) Part. -7 -4 -2 (Adj.) 0 9
The very same, it appeared, two pieces of discolored herring

Order classes, then, can be expected to occur in a fixed interrelationship A string of genitive nouns will show the specified order if they are in one structure; if they do not show the specified order, they are not in one structure. Thus, in a sequence like LESU U5ERBA, (see the paper on Syntagmatic Analysis: Noun-noun Structures) LES is an inanimate concrete noun, like STOL while U5ERB is an inanimate abstract noun such as does not occur in the order sequence determined by inanimate concrete nouns, at least, not in that relative position. Thus, it is possible to conclude that U5ERB is not in a noun-government structure with LES.

Nouns, then, can be coded according to their order categories in relation to specific types of order nuclei. There will be a number of classes of order nuclei such as inanimate concrete, inanimate abstract and so on. There will also be a number of precedence classes which occur with each other nucleus.

Preliminary research also indicates that these and similar classifications will have a number of other uses, besides those for which they are proposed here.

Analysis Based on Transformations

Another path of attack has been explored by Mrs.M. Richman and by Mrs. I. Thompson.

Their study begins with a discussion of the problem of lexical choice on a theoretical level. The structural characteristics of the item which is subject to lexical choice are to be considered; the problem of lexical choice will differ according to whether the item is a function word, a content word, or a mixed word. The associative characteristics of the item are also to be considered;

the general field of discourse of the text is to be taken as more decisive than the association of particular items in the same sentence; where the association of particular items in the sentence is to be considered, each item in the sentence is held to have an 'entry' which introduces it, and an 'exit' after which its role in the sentence is played out, and these entries and exits are the decisive features of the local environment.

The study then discusses the problem of lexical choice on a practical level and describes a particular attempt at a solution. The Russian preposition U was the subject of the research at this level. The GAT concordances, based on about two million running words, were used to collect occurrences of this preposition. Transformations were used in classifying the occurrences, and all those occurrences which were held to be subject to the same transformation were grouped in the same class. Then all the entries and the exits of each transformational class were studied and a search was made for the structural and semantic criteria which determine the classes of entries and the classes of exits. When the structural and semantic criteria were found, a coding was devised for each criterion and the codes were entered under the appropriate entries in the dictionary. The English gloss for the preposition could then be selected on the basis of this coding, and so a type of association (or collocation) coding would be evolved.

Analysis Based on Contrastive Characteristics

Another path of attack is being explored by R. R. Macdonald.

The theoretical point of view is as follows. An acceptable lexical choice can only be made after the structural analysis is complete. The most immediate structural environment is the most decisive in determining the choice. If the immediate structural environment is not decisive, increasingly distant structural relationships are to be probed until a decisive factor is found. The ultimate decisive factor is, of course, the absence of any decisive factor, in which case the most general translation is to be chosen.

1. First, a structural analysis of the source language is to be made. Any item of the source language which is subject to lexical choice will then be seen to have specific structural characteristics. (If there was a structural ambiguity (e.g.; is 'since' a conjunction or a preposition?), the syntactic analysis will have resolved it.) In terms of its structural characteristics, each lexical choice candidate can be described as an immediate constituent of some larger structure, and it is to be investigated first in terms of the other immediate constituent of that structure. The lexical choice candidate is to be contrasted with all other pertinent items of the same form-class which also occur as an immediate constituent in the same structural relationship and in the same or parallel structures.

The meanings assignable to the items in the class of structures so derived are studied, and their similarities and dissimilarities are defined in terms of

the contrasts apparent in the language itself. Each point of similarity of dissimilarity is a component semantic characteristic. These characteristics or groups of these characteristics are then grouped into semantic classes on the basis of their distribution in the language in much the same way as phonetic characteristics or groups of phonetic characteristics are classified into allophones and phonemes.

When the class of structures in which a given lexical choice candidate is one immediate constituent has been exhaustively investigated, the next larger class of structures, in which the already investigated structure class is one immediate constituent, is investigated in the same way. This process of expansion by immediate constituents is continued until the sentence limits a.re reached. (The paragraph and then the entire text are also considered, but not as structural units, which they are not; they are considered as associative units having a specific field of discourse.)

Coding positions are then assigned to each class of semantic characteristic which it has been found necessary to establish, and codes are assigned to designate the individual characteristics in each class; any characteristic or group of characteristics in a class may be isolated from the remainder of the class by a series of dichotomous choices; the coding system is to be capable of unlimited expansion and immediate adaptability.

2. Second, a corresponding analysis of the target language is made. The form class in the target language is analyzed which corresponds to the form class under investigation in the source language. (In some cases it is not necessary *to* analyze every member of the form class in the target language; in other cases there may be no one form class in the target language which corresponds to the particular form class in the source language.)

3. Third, the semantic characteristics of similarity and dissimilarity in the source language (as reflected by the coding) are compared with the characteristics of similarity and dissimilarity in the target language. Where the two systems do not correspond, a transfer system involving further subclassificaton in each language is devised so that there can be a complete correlation between the two systems of semantic classes.

The final coding system, which represents the semantic system of the source language, the semantic system of the target language and a transfer system for a moving from one to the other, is then entered in the dictionary.

Some of the codings are received by the lexical choice candidate itself, and some are received by other items with which it enters into structural relationships. But each choice in each lexical choice candidate has a coding space for every semantic characteristic needed to define it adequately, whether that coding space is filled in the dictionary or not.

Each lexical choice candidate, then, appears in the dictionary with as many sets of coding areas as there are lexical choices. The lexical choice operation

searches the immediate structural environment for an item displaying a code which is entirely or partially the same as that of one of the choices in the range of the lexical choice candidate. If the item displays a code which is entirely similar to that of any choice available in the range of the lexical choice candidate, that particular choice is to be made. If the similarity is only partial, further search for other partials continues. Each cue that is found defines the semantic characteristics of the environment more precisely and will thus narrow down the number of those codings of the lexical choice candidate which are compatible with the environment. Finally one coding remains. That one coding indicates the gloss to be chosen. In some cases, a specific gloss may not be selected immediately; a selection will only be made after glosses have also been selected for the structurally related words, since an acceptable lexical choice may be bound to the specific environment in the target language, and may not be possible under any other circumstances.

As a practical test of this theory, a study of the Russian prepositions is being made. Though the members of the form class of prepositions seem few in number, the variety of their semantic characteristics makes their treatment complicated. The preliminary work is restricted to a limited frame. This frame consists of a preposition as one immediate constituent, and a member of a class of nouns, definable by list, and subsumable under the rubric "nouns of time", as the other immediate constituent.

A certain amount of exploratory research has also been done with the members of another such class of nouns, subsumable under the rubric "nouns of place", as the other immediate constituent with the preposition.

TRANSFER

Transfer is the process of adapting the gloss of each textword to the specific environment in which it will occur in the output.

The transfer mechanisms differ depending on the specific structural characteristics of the input language and of the output language. In the GAT, which is designed to translate from Russian to English, the chief transfer mechanisms are synthesis, insertion and rearrangement.

Synthesis is the process, in the GAT at least, of inflecting or otherwise modifying the canonical form given as the gloss. Verbs, nouns and adjectives are subject to synthesis.

Insertion is the process of introducing words into the English text which do not correspond directly to any word in the Russian text, and so are not given in the dictionary gloss of such a word. Typical inserted forms are the English article, certain English prepositions as transfers of Russian inflection and certain English auxiliary verbs in negative and interrogative structures. 1

Rearrangement is the process of changing the relative order of items or stretches. Typical rearrangement problems are presented by certain type of noun-noun structure, by Russian participles which both are heads of govern-
ment structures and precede the noun head in an agreement structure (nesting and by verb-subject order in Russian declarative sentences.

The succeeding papers of this report are entitled Synthesis, Article Insertion and Rearrangement; each paper deals largely with one particular operation, but cannot be confined strictly to that one because of the manner in which the operations are interdependent and intermingled.

SYNTHESIS

Synthesis is the process of inflecting or otherwise modifying the shape of the glosses of the dictionary entries.

The English glosses in the GAT dictionary are entered in one canonical form. Verbs are entered in the simple non-past form: 'like', 'see', 'go'; nouns are entered in the singular: 'table', 'milk', 'beauty'; adjectives are entered in the positive: 'small', 'good', 'beautiful'. An acceptable translation output will on occasion require that these forms be altered; there must be a method of converting from the simple verb form 'sing', for example, into 'sings', 'singing', 'sang', and 'sung' as necessary, of converting from the singular 'table' to the plural 'tables', and of converting from the positive 'small' to 'smaller' and 'smallest'. This process of conversion from the canonical or dictionary form to other forms is synthesis. The GAT synthesis is, at present, a morphological synthesis only.

Each dictionary entry has space for two glosses. (If further glosses are required, they are listed in ancillary dictionaries.)

The glosses do not necessarily consist of one word; such combinations as 'railroad station' and 'turn out' may appear. In the synthesis of the plural of 'railroad station' the ending is to be added to the second item, 'station'. In the synthesis of the tenses of 'turn out', the various verb endings are to be added to the first item, 'turn'. In short, the point of synthesis differs in different glosses. It is necessary, therefore, to identify the point of synthesis for the computer so that the endings will be added to the item which requires them, and not simply to the last item of the gloss. This identification is achieved by so positioning the glosses in the dictionary that the point of synthesis is always in a predictable location.

The first gloss is entered so that the item to be synthesized ends in position II:80. Thus, in both 'railroad station' and 'turn out', the letter n will stand in that position. Any items of the gloss which follow this point of synthesis are entered in the Overflow Area (III:41-50) beginning at the left. Thus, in 'turn out', the letter o of 'out' will stand in position III:41.

The second gloss is entered so that the item to be synthesized ends in position III:50. There is no provision for overflow in the case of the second gloss; the point of synthesis is necessarily the end of the gloss.

The synthesis pattern of each gloss is indicated by a synthesis code in the dictionary.

Synthesis Codes in General

There is a set of synthesis codes for verbs, a set for nouns, and a set for adjectives and adverbs. It is therefore easiest to describe the codes in three separate groups. However, there are some characteristics which are shared all types of synthesis codes, and these can be described first.

In the GAT synthesis operations, two procedures are used as needed.

a) A number of letters is deleted from an item.

b) A specific group of letters is added to the residue of the item.

These operations are not separately coded; the one coding covers both an instruction as to the number of letters, if any, which are to be deleted, and an instruction as to the identity of the group of the letters, if any, which are to be added. The specific cases given below will exemplify this.

SYNTHESIS

Verbs

From the dictionary form of the English verb (for example: 'sing'), four forms must be synthesized ('sings', 'sang', 'singing', 'sung').

Verb-synthesis Codes

Verb synthesis codes are of two types, depending on whether the first or the second English gloss is selected in lexical choice. The description of the codes for the synthesis of the second gloss must refer to the description of the codes for the first gloss, so that it is necessary to begin the description with the first-gloss codes.

First-gloss codes

The verb synthesis code for the first gloss occupies four positions (I:41-44). Each position receives a symbol, usually alphabetical, which indicates the synthesis of one of the four forms to be synthesized. Each code represents an instruction to delete a specified number of letters, (none, one, or two) from the gloss, and an instruction to add a specific group of letters to the residue.

The First Position (-S-form)

The first position receives a code which indicates how the third person singular of the present is synthesized. There are five possibilities.

Code	Instructions		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
A	0	S	eat	eats
B	0	ES	go	goes
C	1	IES	dry	dries
D	2	S	have	has
E	0	0	must	must

The Second Position (-ED-form)

The second position (I:42) receives a code which indicates how the simple past is synthesized. There are many possibilities if all the irregular verbs are considered; all of the letters of the alphabet and most of the digits are used as codes. The following list gives a selection of the possibilities.

Code	Instructions		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
A	0	D	note	noted
B	0	ED	walk	walked
G	1	IED	try	tried
K	0	TED	chat	chatted
M	0	PED	clap	clapped
P	0	LED	propel	propelled
S	3	OUND	find	found
V	2	AN	run	ran
W	3	OKE	break	broke
Y	2	EW	blow	blew
4	2	PT	keep	kept
5	3	ANK	sink	sank
7	The -ED-form is the same as the -EN-form; see the fourth position (I:44).			

The Third Position (-ING-form)

The third position I:43) receives a code which indicates how the present participle is synthesized. There are nine possibilities.

Code	Instructions		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
A	0	ing	walk	walking
B	1	ing	note	noting
C	0	ting	chat	chatting
D	0	ping	equip	equipping
E	0	ring	bar	barring
F	0	ling	control	controlling
G	2	ying	die	dying
H	0	ning	begin	beginning
I	0	bing	grab	grabbing

The Fourth Position (-EN-form)

The fourth position (I:44) receives a code which indicates how the past participle passive is synthesized. There would be many possibilities if all the irregular verbs that required codings had them. However, only sixteen coding are necessary because of the particular use of the code C.

Code	Instructions		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
A	0	D	note	noted
B	0	ED	walk	walked
C	The -EN-form is the same as the -ED-form (I:42).			

Code	Instructions		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
D	0	N	blow	blown
E	0	NE	go	gone
F	3	SEN	choose	chosen
G	2	AIN	lie	lain
H	1	TEN	write	written
I	2	UN	begin	begun
J	3	OKEN	break	broken
K	0	EN	beat	beaten
L	0	0	bet	bet
M	3	UNK	drink	drunk
N	0	BED	grab	grabbed
0	1	ID	lay	laid
P	3	OLD	sell	sold

The four synthesis code positions are usually cited as a group. Thus, the verb 'work' is described as having the synthesis code ABAB; the verb 'telephone' has the code AABA; the verb 'write' has the code ATBH.

Second-gloss Codes

If the second gloss of a verb is selected during the lexical choice operation, the computer subsequently searches for a synthesis code in I:39. This one-position code is a cover code for that one of the four-position synthesis codes described above which is appropriate to the verb used in the second meaning. The computer generates this four-position synthesis code in I:41-44, replacing any code previously there, and proceeds with the synthesis. Not all irregular verbs are available for second glosses because of the limitations of coding possibilities. It is to be noted that codes E and F have the same value; this is a result of a modification of the original system. The code X in I:73 designates a gloss which has the verb 'be' at the point of synthesis.

Code	Delete	Add
A	0	0
B	old code	X in I:73
C	old code	AHBD
D	old code	AABA
E	X in I:7 3	ABAB
F	X in I:73	ABAB
G	old code	ACBD
H	X in I:73	0
I	old code	ABAD
J	old code	DEBC
K	old code	AKCC
L	old code	BBAB

Code	Delete	Add
M	old code	A2AC
N	old code	A3BL
P	old code	AIBC

Procedure for Verb Synthesis

Verbs in the Russian input are recognized by their parset. The parset is entered in the dictionary or generated by the computer (Positions I:36 ff;). In the unsplit dictionary, each verb form has its parset permanently coded. In the split dictionary, participles have parsets beginning with the codes 23, and all other forms of the verb have the parset 26. In the course of the morphological analysis, infinitives are recognized and their parset is changed to 21; (there is also the alternate (and superrogatory) coding possibility that the parset remains as 26 but that the code I is placed in the nominative-plural location.) Gerunds are recognized and their parset is changed to 24. Finite verbs retain the parset 26 without modification.

The synthesis of each form of the verb will be described separately, following the order of the suboperations in the synthesis operation. Negative (1) and reflexive (2) forms require special prior consideration. Then the active and affirmative forms (3), comprising the infinitive, the present tense, the past tense and the future are dealt with in order. The special case of the verb 'be' (4) is considered separately. Participles and gerunds (5) are treated last.

1. Negative forms

In negative structures, English verb glosses require no change; only the auxiliary verbs which are inserted to accompany them undergo synthesis. The synthesis of these auxiliary verbs is described under the operation for the negative particle, which is lexical choice operation 704 (q. v.).

2. Reflexive forms

In the course of lookup in the split dictionary, the reflexive suffix is removed, and the computer generates the code 3 in I:68.

In the synthesis, if there is the code 3 in I:68, a permanent code in I:40 indicates whether the verb is to be transferred as an active form or as a passive form in English, and the synthesis proceeds on this basis. There are three codings now in use in I:40.

1	Transfer as English passive.
2	Transfer as English active.

5	Transfer using second meaning, which is usually an adjective in this case; perform no further synthesis.

(It will be noted that the codes 3 and 4 are absent. Although these sometimes appear in the dictionary, they are fossils, and are not used in present routines. The code 5 is used only in the routines for the transfer of the two active participles. The code 2 is not actively used to any great extent; a blank has the same effect, and so is now replacing the code 2 in practice.)

If the code 1 is found in I:40, the Russian verb is to be transferred as the English passive. First, the English past participle is synthesized on the basis of the code in I:44. Second, the computer generates the code X in I:73 (BeLoc). This indicates that the English synthesis requires the insertion of the verb 'be'. The operation for the synthesis of the verb 'be' is described later in this paper.

3. Active and Affirmative forms

The computer searches the parset to determine whether the verb is in the infinitive, the past, the non-past, or the future, (with BUDU, BUDET, etc.). There are four suboperations provided for these four possibilities.

A. Infinitive

The computer inserts 'to' in EPREP (II:36-44) unless certain conditions are present.

a) If there is a preceding Russian verb which has the code 4 in I:68, 'to' is not inserted. Verbs which have this code are those which are glossed with English verbs which are followed by other verbs without the infinitive marker 'to' (can, let, must, etc.).

b) If there is a preceding Russian verb whose English gloss is a verb which is followed by a verbal noun in '-ing' rather than by an infinitive, 'to' is not inserted and the gloss receives the suffix '-ing' according to the code in I:73. The Russian verbs in question are those which are glossed by 'avoid', 'detest' and similar verbs.

c) If the word ESLI: 'if', precedes, the words 'it is' are inserted and the infinitive is changed to the past participle according to the code in I:44.

B. Non-past tense

The dictionary listing is used as the translation, without synthesis, unless the textword is coded as third person singular. In this case, the computer synthesizes a form in accordance with the code in I:41.

C. Past tense

The computer synthesizes *a* form in accordance with the code in I:42.

4. The 'be' Routine

For purposes of machine translation, the synthesis of all forms of the verb 'to be' is the same whether 'to be' functions as a main verb or as an auxiliary.

A. Infinitive

If the Russian verb which has 'be' at the point of synthesis of its gloss is an infinitive, the routine is the same as in the case of other verbs.

B. Present Tense

If the Russian verb which has 'be' at the point of synthesis of its gloss is in the third person singular, 'is' is used as the translation. If the verb is in the first person singular, 'am' is used as the translation. Otherwise, 'are' is used as the translation.

1

C. Past Tense

If the Russian verb which has 'be' at the point of synthesis of its gloss is in the first or third person singular of the past tense, 'was' is used as the translation. If it is in the second person singular or in the plural of the past tense, 'were' is used as the translation.

5. Participles and Gerunds

The negative-particle operation does not handle the translation of NE before participles; where NE precedes a Russian participle, 'not' is inserted immediately before the English participle.

A. Past Passive Participle (Parset 231)

The English past passive participle is synthesized in accordance with the code in I:44. If the participle is to the left of its noun head, no rearrangement or insertion is necessary; if it is to the right of its noun head, 'which was' is inserted if the participle is singular and 'which were' is inserted if the participle is plural.

B. Past Active Participle (Parset 232)

For the transfer of the Russian present active participle, three possibilities are to be considered.

a) If the participle has the reflexive suffix (code 3 in I:68) and also has the code 5 in I:40, the second meaning is chosen. If the second meaning is an adjective, as it usually is in this case, no further synthesis or rearrangement is necessary.

b) Otherwise, if the participle is to the left of its noun head, the simple English past tense is synthesized according to the code in I:42; the participle is rearranged so as to stand at the right of the noun head; the word 'which' is inserted to the left of the participle.

c) If the participle is to the right of its noun head, the simple English past tense is synthesized, and the word 'which' is inserted to the left of the participle.

C. Present Passive Participle (Parset 233)

The Russian present passive participle is transferred into English as the English past passive participle, which is synthesized in accordance with the code in I:44. If the participle is to the left of its noun head, no rearrangement or insertion is necessary; if it is to the right of its noun head, and if it governs a nominal form, 'being' is inserted before the English past passive participle.

D. Present Active Participle (Parset 234)

For the transfer of the Russian present active participle, five possibilities are to be considered.

a) If the participle has a reflexive suffix (code 3 in I:68) and if it also has the code 5 in I:40, the second meaning is chosen; if the second meaning is an adjective, as it usually is in this case, no further synthesis or rearrangement is necessary.

b) If the participle has a reflexive suffix (code 3 in I:68), but does not have the code 5 in I:40, the English past passive participle is synthesized according to the code in I:44. If the participle is to the left of its noun head, it is rearranged so that it stands to the right of the noun head. The words 'which is' are inserted if the Russian participle is',' singular, and the words 'which are' if the Russian participle is plural.

c) If the participle has no reflexive suffix and is to the left of the noun head, the English present participle is synthesized according to the codes in I:43. Rearrangement is not necessary.

d) If the participle has no reflexive suffix, and is not preceded by the negative particle NE, and is located to the right of the noun head, the English present tense is synthesized according to the code in I:41, and the word 'which' is inserted. Rearrangement is not necessary.

e) If the participle has no reflexive suffix, and is preceded by the negative particle NE, and is located to the right of the noun it agrees with, the English gloss is used without synthesis, the particle NE is transferred as zero, and the words 'which do not' are inserted if the Russian participle is plural.

Gerunds, outside of a few exceptions specially glossed in the dictionary or in lexical choice, are to be transferred by the synthesis of the English -ING- form, according to the codes in I:43.

SYNTHESIS

Nouns

Russian nouns exhibit inflections which signal four structural categories: case, number, gender and animateness.

These four categories are not equally as important in English structure as in Russian structure, nor are they always signaled by the same structural mechanism of inflection.

Case

The effect conveyed in Russian by the category of case in the inflection of the nouns is conveyed in English by prepositions or by features of word order. The transfer of Russian case inflections in the GAT is handled by special operations similar to the lexical choice operations. Here are samples of such operations.

(The description of these operations mentions 'special lists'.

(A Russian word commonly determines a particular case inflection in a government structure in Russian. Its English gloss may require a preposition (or zero) which is not a generally acceptable gloss for that case inflection. Thus, the Russian verb JDAT6: 'wait', determines the genitive inflection in its object; the generally acceptable glosses for the genitive inflection are 'of' and zero. The English gloss 'wait' requires the preposition 'for'; neither 'of' nor zero is acceptable in such a sentence as 4 JDU OTVETA: 'I am waiting for an answer'. Therefore JDAT6 is specially listed as determining a genitive inflection which is best transferred by the preposition 'for'. Similarly, the Russian verb POMOC6: 'help', determines the dative inflection in its object; the generally acceptable gloss for the dative inflection, after verbs, is 'to'. The English gloss 'help' requires no preposition; ONA POMOJET MATERI: 'She will help her mother'. Therefore POMOC6 is specially listed as a verb which determines a dative inflection that is best transferred as zero.

(None of these special lists is particularly long.

(A suggestion has been made for replacing the lists by a simpler operation. This will appear as a part of a projected paper on English Synthesis.)

Genitive Inflection Transfer	Y	N	A
1. Let the noun in the genitive case be i.	-	-	2
2. Does i have the stretch code 3112 (agreement stretch)?	3	5	–
3. Count the number (x) of items to the left of i in the 3112 stretch.	-	-	4

Genitive Inflection Transfer	Y	N	A
4. Let a be x +1.		-	
5. Let a be 1.		-	
6. Is i a sequent in a government stretch ?	7	9	
7. Is the head of the government stretch on a special list?	10	8	
8. Is i the first noun in the stretch?	9	11	
9. Transfer the genitive inflection as zero.	-	-	9
10. Transfer the genitive inflection by making the insertion that the special list calls for in EPREP of i-a+1.	-	-	9
11. Transfer the genitive inflection by inserting 'of in EPREP of i-a+1.	-	-	9

Dative Inflection Transfer	Y	N	A
1. Let the noun in the dative case be i.			1
2. Does i have the stretch code 3113 (agreement stretch)?	3	5	
3. Count the number (x) of items to the left of i in the 3113 stretch.	-	-	4
4. Let a be x+1.	-	-	6
5. Let a be 1.	-	-	6
6. Is i a sequent in a government stretch?	7	8	
7. 8 -Is the head of the stretch on a special list?	9	10	-
8. Transfer the dative inflection as zero.	-		99
9. Transfer the dative inflection by making the insertion that the special list calls for in EPREP of i-a+1.	-	-	99
10. Transfer the dative inflection by inserting 'to' in EPREP of i-a+1. .	-	-	99

Number

The effect conveyed in Russian by the category of number in the inflection of the nouns is conveyed by inflection in English also.

Thus, English nouns require to be synthesized for number.

If a Russian noun which has both singular and plural forms is in the plural, its English gloss is synthesized as a plural. The noun synthesis operation is similar to the verb synthesis operation. The codes are of the same type, but have, of course, different values.

Noun-Synthesis Codes

The Noun-synthesis codes are located in position I:72 of the dictionary. This location is named NounS. (It is also named AdjR, since the adjective synthesis codes are also located here.)

Code	Instruction		Examples	
	Delete	Add	Dictionary Form	Synthesized Form
A	0	S	house	houses
B	0	ES	pass	passes
C	1	IES	fly	flies
D	2	ES	basis	bases
E	0	0	sheep	sheep
F	2	A	maximum	maxima
C	1	VES	leaf	leaves
H	2	I	radius	radii
J	0	E	alga	algae
L	4	EETH	tooth	teeth
M	4	ICE	mouse	mice
P	2	VES	life	lives
Q	2	EN	man	men

Gender and Animateness

The effect conveyed in Russian by the categories of gender and animateness in the inflection of the noun are only approximately matched in English nouns by the category of gender. English gender is generally signaled by the identity of the noun itself. In the scientific texts on which the GAT has been tested, the category of gender is relatively unimportant, since almost all the nouns encountered are neuter.

A rough approximation serves the immediate purpose of transfer. By this approximation of the Russian and English categories, the Russian masculine animate noun is taken as masculine in English, the Russian feminine animate noun is taken as feminine in English, and all other nouns are taken as neuter in English. As texts in other fields come to be used, however, the English glosses will have to be coded for gender (masculine, feminine, common, neuter and personified neuter) independently of the Russian.

SYNTHESIS

Adjectives

Russian adjectives exhibit inflections which signal four structural categories: case, number, gender and animateness. They also exhibit, along with adverbs, variations in form which signal degrees of comparison; in some, these variations are effected by the selection of certain suffixes: in others, the variations are effected by the collocation of specific words immediately to the left,

The effect conveyed in Russian by the categories of case, number, gender and animateness in the inflection of the adjectives is that of indicating agreement with some nominal form which is, if present, in the same case, number, gender and animateness category. This effect in English is conveyed by the position of the adjective in relation to the nominal or to its surrogate. The transfer of these adjective inflections, then, is best achieved by rearrangement, in the few cases where rearrangement is necessary.

The effect conveyed in Russian by the suffixation and collocation which designates degrees of comparison is best transferred in English by suffixation and collocation also, though the distribution of these features in English is far different from that in Russian.

If a Russian adjective or adverb is in the comparative or the superlative form, the English adjective or adverb which serves as its gloss may undergo synthesis. The adjective synthesis operation is similar to the verb or noun synthesis operation. The codes are slightly different in type and have, of course, different values.

The synthesis code for a noun or for a verb represents an instruction to delete a specified number of letters from the item at the point of synthesis, and an instruction to add a specific set of letters to the residue. The synthesis code for an adjective or an adverb, however, refers to two specific sets of letters, either of which may be added to the residue. The one is to be added if the comparative form of the adjective or adverb is required. The other is to be added if the superlative is required. If the adjective or adverb is not in itself marked as comparative or superlative, the immediately preceding text item is checked. If the item is BOLEE, the adjective or adverb is read as comparative. If the item is NA1BOLEE or a compatible form of SAMY1, the adjective or adverb is read as superlative.

If the English gloss has no code in I:72, its comparative is formed by inserting the word 'more' in EPREP, and its superlative by inserting 'most'. If the English gloss has a code in I:72, the instruction covered by that code is carried out, and the ending appropriate to the comparative or superlative form is chosen.

Adjective-adverb Synthesis Codes

The adjective-adverb synthesis codes are located in position I:72 (AdjR). The following are a selection of these codes.

Code	Instruction			Examples		
	Delete	Insert		Dictionary Form	Synthesized Form	
A	0	R	ST	pure	purer	purest
B	0	ER	EST	high	higher	highest
C	0	BER	BEST	drab	drabber	drabbest
D	0	DER	DEST	red	redder	reddest
E	0	GER	GEST	big	bigger	biggest
F	0	MER	MEST	grim	grimmer	grimmest
C	0	NER	NEST	thin	thinner	thinnest
J	0	TER	TEST	hot	hotter	hottest
K	1	IER	IEST	happy	happier	happiest

Article Insertion

There is no single structural class in Russian which comes close to conveying the effect conveyed by the definite and indefinite articles in English The features of Russian which do convey this effect are of all types: morphological, syntagmatic, syntactic, and semantic. In some cases, perhaps, the hearer must adduce the information from the circumstances and not from the language at all.

For this reason, the insertion of the English article has remained a problem.

The solutions that have been suggested are successful as far as they reach. But none of them reaches far. They are effective, unfortunately, only in those cases where the presence or the identity of the article is of less importance. Those cases where a specific article is necessary to an exact translation constitute a problem which still evades solution.

The first attempt to solve the problem of the articles for the GAT was made by Philip H. Smith. The program produced an output of variable quality A rather involved structure might be translated into English with each of the articles in impeccable place. A rather straightforward structure might prove not to be acceptable in translation. Some of the difficulties are assignable to such factors as coding or programming, but by far the largest share of the difficulties arise from the problem of transforming the vaguest and most heterogeneous of Russian indications into three well-marked but till now only nebulously described patterns in English. In addition, the analysis was based on a limited corpus from a text in chemistry and this bias is reflected in the analysis to a considerable degree. At the end of his description of the analysis, Smith wrote:

> "Indeed, one may question whether article insertion, at the present stage of machine translation, is justified at all. If the writer could claim 90% "satisfactory" article insertion (to be determined, say, by a panel of native English speakers) for his routine, he would be very happy. But even at the high level of 90%, who will be satisfied with 10% actually wrong and misleading article insertions? At present we must say that article insertion is little more than an exercise, a stunt perhaps, and that, while extremely interesting, it serves a doubtful purpose. It already seems probable that article insertion cuts across all levels of analysis, from the simplest syntagmatic arrangements to the subtlest semantic situation. Further research on article insertion will doubtless provide interesting insights into the operation of language on all levels, and will surely be a useful activity in machine translation. In the meantime proper article insertion requires, at least, perfect structural analysis of the source language, and beyond that it needs large quantities of semantic information to operate properly. "

A sketch of Smith's Article Insertion Operation will give an indication of the method.

Russian nouns, adjectives, adverbs, and verbs are classified according to the relationships of their glosses to the definite and indefinite articles.

The threefold choice among the occurrence of the definite article, the occurrence of the indefinite article and the occurrence of no article is first reduced to a series of binary choices.

The occurrence of no article is accepted as a special case of the occurrence of the indefinite article. Accordingly, the first binary choice is that between the definite article and a non-definite article. If the non-definite article is chosen, the second binary choice is that between the indefinite article and no article; this choice is almost entirely a function of the individual noun in question.

Russian nouns, adjectives, adverbs and verbs are assigned to article-insertion classes according to the relationship of their glosses to the definite and indefinite articles. There are thirteen classes of nouns, four of adjectives, nine of verbs, and one of adverbs.

The operation for article insertion in relation to verbs is quite separate from that for article insertion in relation to nouns, adjectives and adverbs. Since the latter operation is the more involved, it serves as the better example.

The noun construct which principally requires the insertion of the definite article is the head of a noun-government stretch in which the dependent is a noun in the genitive. (Such a stretch has the stretch code 1122 in the GAT system, and may be referred to, for the sake of brevity, as a 1122 stretch,)

It is a general rule that the noun which governs another noun in the genitive requires the definite article before its gloss. A few special cases must be given precedence over this general rule, and, of course, a few exceptions to it must be kept in mind.

A noun construct consists of an agreement stretch (a noun and adjective) and of any modification stretch (an adverb and adjective) which depends on that government stretch. Consequently, each noun, adjective, or adverb which belongs to an article-insertion class if important to this operation.

Article Insertion Codes

Every dictionary entry which is a member of an article-insertion class receives an article-insertion code. The article-insertion codes are stored in position I:79. They are subdivided according to parts of speech.

Article-Insertion Codes for Nouns

Code	Class Represented
9	This class includes proper names, the names of cities and months, but not the names of rivers. No article is inserted with such a noun.
A	This class contains the one word ATMOSFERA. The indefinite article is inserted when ATMOSFERA governs the genitive, otherwise the definite article is inserted.
D	Examples of this class are AVTOR and DNO. The definite article is inserted unless prior considerations prevent this.
G	When a member of this large class is the head of a 1122 stretch, the definite article is inserted before the dependent, unless prior considerations prevent this.
M	This class contains the one item MESTO, and that only when lexical choice has selected the gloss 'point'. When MESTO: 'point' is the head of a 1122 stretch, no article occurs before the dependent.
O	Nouns of this class are the mass nouns ('water', 'chlorine', 'sugar') and those "process" nouns (deverbals, e.g. 'precipitation') which are not included under P. The indefinite article is never inserted before these nouns.
P	This class contains process nouns. The indefinite article never occurs before these nouns. When such a noun is the head of a 1122 stretch, the definite article is inserted before the dependent.
Q	This class contains SOL6, SPIRT, and 3FIR, each of which can represent a single chemical compound or a class of compounds. This class was formed in the belief that these nouns behaved differently from those of the class coded V. The belief was ill-founded; Class V and Class Q nouns behave alike.
S	This class contains QVET: 'color'. When QVET is the dependent in a 1122 stretch, the definite article does not occur with the head.
T	This class contains the item TIP: 'type'. If TIP is either the head or the dependent in a 1122 stretch, the definite article does not occur with the head.
V	This class contains nouns from the field of chemistry which, when alone, represent a class of compounds but which, when modified, may represent a specific chemical compound. Thus KISLOTA: 'acid' alone names a class of compounds; there is no restriction on the insertion of an article before its gloss. But SERNA4 KISLOTA:

'sulfuric acid', refers to a unique chemical compound; the indefinite article is not inserted before its gloss. SILNA4 KISLOTA: 'strong acid', follows the pattern of KISLOTA alone, however.) Thus, the insertion of the indefinite article is conditioned by the identify of the adjective, if any, which qualifies the noun.

X This class contains nouns which govern a partitive genitive (an atom of chlorine, a drop of water). When such a noun is the head of a 1122 stretch, the definite article is not inserted before its gloss on that account although it may be inserted for other reasons.

This class contains SKOROST6: 'rate'. When this noun is the head of a 1122 stretch, no article is inserted before the dependent unless the dependent is qualified by an adjective, in which case the definite article is inserted. ('The rate of reaction' as against 'the rate of the oxidation reaction'.)

Article-Insertion Codes for Adjectives

Codes	Class
A	This class includes such adjectives as ZNACITEL6NY1: 'considerable' and IZVESTNY1: 'certain'. The indefinite article is inserted before the noun construct, unless the noun is so coded as to prevent this.
D	This class includes the ordinal numerals and other adjectives such as VYWEOZNACENNY1: 'above-mentioned'. The definite article is inserted before the nominal construct.
O	This class includes such adjectives as SERNY1: 'sulphuric'. The indefinite article is not inserted before the nominal construct when the noun has the code V or the code Q.
9	This class includes MO1: 'my', and 3TOT: 'this'. No article is inserted before the nominal construct.

Article-Insertion Codes for Adverbs

Codes	Class
D	This class includes SAMY1: 'most'. The definite article is inserted before the nominal construct.

The Article-Insertion Operation

A The article-insertion operation checks the absolute cases first.

a) Any nominal construct which contains an item having the code 9 is marked as receiving no article.

b) Any remaining nominal construct which contains an item having the code D receives the definite article.

B The operation checks the noun-noun government stretches (1122) next.

c) Any remaining nominal construct which is the head of a 1122 stretch receives the definite article unless the head has the code A, S, T, or X or unless the dependent has the code T.

d) Any remaining nominal construct which is the dependent of a 1122 stretch receives the definite article if the head has the code G or the code M or if the dependent is not qualified by any adjective and the head has the code Z.

C The operation completes the task of inserting definite articles.

e) Any remaining nominal construct in which the noun has the code A receives the definite article.

D The operation eliminates the cases where no article is to be inserted.

f) Any remaining nominal construct in which the noun has the code P, or in which the noun has the code V or Q and an adjective has the code Q, is marked as receiving no article.

g) Any remaining nominal construct receives the indefinite article.

A cursory reading of the codes will show that this operation is closely bound to the chemical corpus and that it is ad hoc to a very great degree.

This operation is now in disuse.

A subsequent attempt to solve the problem of article insertion was made by Mrs. M. Richman. The general outline of this analysis was similar to that of Smith's, but Mrs. Richman worked with a larger and more varied corpus, and consequently, she produced a much more intensive analysis of all kinds of factors, especially of semantic factors.

This analysis has not yet been programmed, and so has not been tested in action, in part because it would require extensive changes in

dictionary coding, and in part because many of those who work with article insertion are asking the same question that Smith originally asked: Is it better to insert the articles with the assurance that a certain percentage of them will be unacceptable, or to omit them and so to produce a text in telegraphic style?

At the moment, greater consideration is being given to the latter alternative. The scientific texts which are being translated now are understandable without articles. In addition, experiments have been tried with various texts from which the articles had been eliminated, and these tend to show that the native speaker of English has little difficulty reading such texts, that the ambiguities are few, and that there is very little disagreement among native speakers as to how each would reinsert the articles in such a text. Thus the insertion of the article begins to seem a luxury rather than a necessity.

Nonetheless, general research on article insertion is being continued by L. E. Dostert.

R. R. Macdonald is elaborating a system of cues for article insertion from the occurrences of other prepositions than 'of'. These cues are valuable only in prepositional phrases; this tends to substantiate the idea that article insertion is variously connected with all levels of the structure of the language.

Rearrangement

Rearrangement is necessary at many levels of the GAT system. Word-by-word translation fails to solve a wide range of problems. The crux is that, in Russian, inflection is of primary importance and word order is secondary, while in English, word order is of primary importance and inflection is secondary. Russian syntactic units must first be recognized by their inflection they must then be transferred at various levels and in certain orders depending on their character; they must finally be rearranged in the rather strict English order.

A few rearrangement problems are solved by the idiom routine (BROMISTOGO VODORODA becomes 'hydrogen bromide'.) Other problems are solved by the nesting routine; almost every free nesting involves a change in the order of the textwords. (BOGATYE METANOM GAZY becomes 'gases rich in methane',) Still other problems are solved by the participle synthesis routine. (UMEN6WIVWA4S4 TEMPERATURA becomes 'the temperature which decreased'.)

The largest problems, however, are those which involve the rearrangement of the output of the syntactic routine. The relatively free order of the subject and predicate in Russian must be altered to the relatively fixed order of the subject and predicate in English. The greatest difficulty is in those cases where the Russian order is: object-verb-subject, and where the relationship is shown by the inflections of the subject and object. Since there are no corresponding inflections in English (the relationship being shown by the order of the items), transfer without rearrangement produces the ambiguous effect that the object is the subject and that the subject is the object. Thus a specific Russian sentence transferred without rearrangement produces the translation

'Further study prevented clouding of the liquid',

whereas transfer with appropriate rearrangement produces the more acceptable translation

'Clouding of the liquid prevented further study'.

The following examples show problems of rearrangement at all levels.

Russian Order	English Order
Carbonate copper	Copper carbonate
Here not occurs	Does not occur here
Upon the action of bromine on the isolated by us spirocyclodecane	Upon the action of bromine on the spirocyclodecane isolated by us.

Russian Order	English Order
The action on the chlorohydrins of alkali	The action of alkali on the the chlorohydrins
Vulcanization its	Its vulcanization
Established was the tendency of the halides of mercury to complex formation with the halides of alkali metals and ammonium in the metals	The tendency of the halides of mercury to complex formation with the halides of alkali metals and ammonium in the metals was established
Further study prevented clouding of the liquid.	Clouding of the liquid prevented further study.

The study of these and similar cases has shown that fifteen patterns of rearrangement are needed. These are:

Russian Order	English Order
1 a b	b a
2. a...b	a b
3. a b c	b a c
4. a b c	c b a
5. a b c	a c b
6. a b c	b c a
7. a b c	c a b
8. a . . . b c	a c b
9. a b . . . c	a..b c
10. a b c	a c b c
11. a b c d	a d b c
12. a b c d	a c b d
13. a b c d	a c d b
14. a b c d	b a c d
15. a..b c d	a c b d

These patterns combine in various ways. The total number of combinations which may exist is enormous. Fortunately, the number which actually exists is rather small. In addition, it is possible to reduce the complexity of the pattern combinations by dealing with the rearrangement patterns in three steps.

The first step deals with the rearrangement of the subject-complex and the predicate-complex relative to each other.

The second step deals with the rearrangement of items inside each of these two types of complex.

The third step deals with the further rearrangement between the subject and predicate complexes which has become necessary because of the internal rearrangement dealt with in the second step.

This procedure can be completely described by the formulation of fifty-two generalized statements. Those dealing with the rearrangement of syntagmatic and syntactic stretches are numbered Q-01 to Q-32. Those dealing with the rearrangement of individual words are numbered R-01 to R-20.

These generalized statements are of the type of:

Q-01 In a 1:0 type sentence, if a 1122 stretch is interrupted by a 512X stretch, place the 512X stretch after the second part of the 1122 stretch and join the two parts of the 1122 stretch (pattern 5)

Q-30 In a 0:1 type sentence, if a noun which is in the accusative case and which has the code 311X has been designated the subject during the course of the syntax routine, place the subject and its 311X stretch before the predicate. (Pattern 1).

Q-32 In a 2:1 type sentence, if the predicate carries the code 2123 or 2124 or 2520 and is followed by the subject, place the subject before the predicate* (Pattern 1).

R-01 When an adverb coded 1 in position I:80 is between a participle and a noun, and the participle does not carry the code 311X, place the adverb before the participle. (Pattern 3).

R-18 In a predicate complex, if the predicate word is not preceded by an adverb a conjunction, a comma or a preposition coded 512X, place the predicate word in the first position of the predicate complex. (Pattern 1).

Rearrangement Codes

The rearrangement operation requires very few specific codes of its own.

Only certain adverbs are assigned codes which are intended specifically for use in the rearrangement operation. These codes are entered in position I:80, and are essentially identification codes which enable certain adverbs to be recognized more quickly than by the process of matching the words themselves.

Apart from these, the rearrangement operation utilizes codes associated with the Russian textwords and with the Russian analysis. When the synthesis of the English gloss is completed and any necessary insertions have been made, the rearrangement operation scans the Russian text for indications of the rearrangement processes to be effected and proceeds to apply them to the glosses. This rearrangement of the glosses is then printed out as the English translation.

At present, each of the fifty-two generalized statements has its own routine. These routines are readily derivable from the statements as given above.

For example, statement Q-32 may be summarized as follows:

Q-32	Y	N	A
1. Is the sentence a sentence of syntactic type 2:1?	2	99	-
2. Is the predicate a member of a 2123, 2124, or 2520 stretch?	3	99	-
3. Does any subject follow the predicate?	4	99	
4. Call the predicate a.	-	-	5
5. Call the subject or subjects that follow the predicate b.	-	-	6
6. Rearrange according to pattern 1.	-	-	99

If further research shows that considerably more than fifty-two generalized statements are needed, it will be advisable to program the fifteen rearrangement patterns individually. Then the present fifty-two generalized statements can be resolved into their component patterns and dealt with one pattern at a time. Any new generalized statement can be treated in the same manner. In this way only fifteen operations need be used, but more than one may be used for each statement. This will undoubtedly prove more economical than devising new operations for a rapidly increasing number of general statements.

The routines for the fifteen operations have already been prepared, and can be used as necessary.

While the rearrangement operation is strong in theory, the routine is rather weak in practice. This is not due to any intrinsic features of the routine itself, but to the circumstance that it is the last routine and so is dependent on all that goes before it. A slight flaw in the morphology, in the syntagmatic analysis, or, most especially, in the syntactic analysis can completely disrupt the rearrangement routine output.

Detailed work on the rearrangement operation must wait until the elaboration of the previous operations is further advanced. The rearrangement routine has not been used recently in producing output.

Further Research: Self-organization

The transfer operations described here are almost entirely oriented to the Russian text. Thus, an English gloss is considered to be plural if the Russian textword is plural. English prepositions are inserted on the basis of Russian case forms, or on the basis of special codings of the Russian textword. The syntactic rearrangement is effected on the basis of information derivable entirely from the Russian input or from the Russian analysis.

The suggestion has been made that the transfer would be more acceptable in its finer points if the synthesis, insertion and rearrangement were linked to the identity of the English glosses selected rather than to the identity of the Russian textword.

Thus, if a Russian entry has a variety of English glosses, each gloss will receive its own synthesis codings; this will eliminate the restrictions as to synthesis and overflow at present imposed on the second gloss and can also eliminate the restriction to two glosses in the main dictionary.

Again, an English gloss which determines the occurrence of a specific preposition ('depend on', 'consist of), would receive coding to permit of inserting that preposition without recourse to the special lists which must now be stored to handle lexical choice and case inflection transfer.

Similarly, an English gloss which determines a particular structural sequence which is not signalled by the Russian text ('refuse to go', but 'detest going'; 'I like to' instead of 'I like') would receive a coding appropriate to the particular English gloss and not to the Russian textword.

In this way, the English synthesis need no longer be restricted to the morphological level, but expanded to include syntagmatic, syntactic and semantic features as well. The process of translation would become less a transfer from Russian into English with the attendant possibility that the English will be affected to a large extent by the peculiarities of the Russian, than an analysis of the Russian to extract all possible information followed by a synthesis of English in a pattern of self-organization so as to preserve all of the information, but to frame it in a purely English setting.

One important product of such a system would be the evolution of an English synthesis operation which would be more independent of the Russian analysis, and so more generally useful in other machine translation systems. Indeed, the English analysis which is part of the English-to-Turkish Translation System has been suggested as the basis of a self-organizing transfer to English, whether from Russian, or from any other language.

COMPARATIVE MACHINE TRANSLATION ANALYSIS IN SLAVIC

The comparative machine translation analysis of four Slavic languages was conducted to ascertain whether the similarities in structure of a group of related languages might permit of developing a common system of syntax which would facilitate machine translation to and from those languages. The research might also indicate whether such a common system of syntax is useful for groups of languages which are not related.

The possibility of a common general syntax for a set of related languages was suggested by Professor L. E. Dostert in his paper, "An Experiment in Mechanical Translation; Aspects of General Problems", which was delivered before the Scientific Literature Section of the American Chemical Society in September 1954.

In Professor Dostert's suggestion, the syntax of each member of a set of natural languages was to be convertible into a core syntax, and the core syntax was to be convertible into each natural syntax. Such an arrangement would permit of treating each of the natural languages as either source or target; it would also permit of achieving translation with a minimum of analytical work, since each language need be analyzed only in relation to the core, and not in relation to every other language of the set.

The Slavic languages recommend themselves for comparative research because of their very close relationship. Accordingly, when it was decided to embark upon comparative research at Georgetown, Professor Dostert's suggestion was accepted as a basis and the research focused on four of the Slavic languages: Russian, Czech, Polish and Serbo-Croatian. These languages represent the different branches of Slavic (Russian: East Slavic; Polish and Czech: West Slavic; Serbo-Croatian: South Slavic). Moreover, staff was immediately available to do research in these languages.

Method of Procedure

It was agreed that the comparative machine translation research would be conducted on the graphic, morphological, syntactic and semantic levels.

At each level, the task to be performed was the comparison of the forms of the four languages and the classification of these forms in terms of absolute equivalences, partial similarities and absolute differences.

An absolute equivalence exists where the linguistic function of the forms is identical in all four languages. A partial similarity exists where the function of the forms is identical in some but not all of the languages. An absolute difference exists where the linguistic function of the forms is different in all four languages.

The pattern of equivalences, similarities and differences was to be noted at each level of analysis, and similarities were to be classified as to whether they occur in two, three, or four of the languages of the set.

The common syntax was to be designed to include programs which could handle the equivalences and the most frequent types of similarities.

It was anticipated that the syntactic similarities would be more significant than the morphological similarities, and that it would be necessary, in the long run, to analyze each language as both target and source language in order to provide a useful common system.

The final aim was to be the discovery of the types of configuration which produce the greatest efficiency in the transfer algorithms.

The preliminary research was primarily in the field of morphology. The results of this research are given below.

The Transliteration System

The proposed system for the transliteration of the Cyrillic or Latin characters in the Russian, Czech, Polish and Serbo-Croatian alphabets does not require a detailed knowledge of the orthographies or of the techniques of phonological analysis. The transliteration is computer-oriented and permits of mechanical reconversion to the original alphabets. The use of digraphs conduces to the legibility of the transliterated text, and this was considered more important than the economy of space possible with a transliteration system which uses only single symbols for each letter of the original orthography.

The total number of different Cyrillic and Latin characters in the four languages is eighty. These are represented in the transliteration by fifty-one signs, of which twenty-five are composed of single symbols, and twenty-six are digraphs composed of one of the single symbols combined with one of the special symbols used only in digraphs.

I

The Morphological Analysis of Polish Nouns

The morphological analysis of Polish nouns for machine translation purposes aims at the establishment of mutually exclusive classes. Polish is treated only as a source language in this particular analysis. The analysis will have to be somewhat modified, therefore, if Polish is later to be used as a target language.

The analysis was based on a transliteration of the orthographic forms, and on a selection of about 8500 commonly used words.

The first step in the analysis was the segmentation of noun forms into base and ending.

Each base was classified according to its pattern of distribution in relation to the other bases and to the inflectional endings. The bases were assigned to different classes if they are in contrast and to the same class if they are in complementary distribution. Unique base alternants, which occur with one suffix only, are listed in the full-form dictionary, and are not split into base and ending.

A total of fifty-four classes of bases was established: thirty-four can be characterized as masculine, eleven as feminine, six as neuter and three as plural. Most classes contain from three to five subclasses. Alpha-numerical identification codes were attached to the classes and subclasses, and these classes and subclasses, along with both their identification codes and the endings with which they occur, have been organized into tables for convenience.

The inflectional endings of Polish mark the grammatical categories of gender, case, number and animateness. (Animateness is marked on the morphological level, in masculine nouns only.) The analysis recognized thirty-eight endings; these were divided into five subclasses on the basis of their length in terms of the number of transliteration symbols; they were grouped into twenty "morphemes" on the basis of their distribution.

The morphological functions of the endings are shown in tables in which the cases are listed vertically and the classes of bases horizontally. Such a table shows, on the one hand, the morphological function of a particular ending when associated with a certain class of base, and, on the other hand, the classes of bases in which a particular case is marked by the ending in question. The morphological function of each ending in relation to those base classes with which it occurs is shown in a separate table, and those endings which have more than one morphological function are distinguished from those with one only.

Four patterns of morphological value were established:

1. unambiguous,
2. ambiguous in case, unambiguous in number (singular only),
3. ambiguous in case, unambiguous in number (plural only), and
4. ambiguous in both case and number.

The classification system for Polish nouns is an open system; if additions prove necessary later, they can be made without the need for changes in the logical system.

The Morphological Analysis of Polish Verbs

The analysis into bases and endings was based on the transliteration of the orthographic forms, rather than on strictly morphemic or phonemic boundaries.

The first step in the analysis was a random selection of several hundred verbs and the listing of their paradigms. The most frequently recurring endings were split off. Only one split was made in any form, and the location of the split was determined by a different criterion in each approach.

Three different approaches were followed.

In one approach, the split between base and endings was located on the basis of morphophonemic alternations in the verb form. The split was so placed that the alternations were included in the base rather than in the ending. By this procedure it was possible to keep the total number of suffixes at a minimum while the number of bases per verb was expanded.

In a second approach, the split between base and ending was located in such a way that the alternations were in most cases included in the ending. this procedure the number of suffixes was increased and the number of bases per verb was reduced.

The remaining approach differed from the first two. In this approach, the endings were analyzed and segmented into components according to their grammat'ical value. This approach is not considered practical for machine translation at the present time, but it is of value for teaching purposes because it provides the student of Polish with a clear picture of the structure of verbal suffixes.

In addition to the three approaches mentioned above, a fourth was also considered; this was the classification of verbs on the basis of base alternations. After a short investigation it was found that the number of base alternations is so high in Polish that such an analysis would be too complicated if it included all alternations. However, it would be profitable to include certain frequent base alternations because this procedure would decrease the number of bases listed as separate entries in the dictionary. Such an analysis could be used either alone or in conjunction with one of the approaches described above.

The Morphological Analysis of Adjectivals in Czech and Serbo-Croatian

The morphological analysis of adjectivals in Czech was based on the same principles as were the other analyses described in this report.

In addition to the analysis of inflectional morphemes, an analysis of the derivational morphemes which mark the comparative form was carried out, and a procedure for the automatic recognition of the comparative form was outlined.

The same technique of analysis was also applied to Serbo-Croatian adjectives.

The Morphological Analysis of Serbo-Croatian Verbs

The data for the morphological analysis of Serbo-Croatian verbs was collected by a native speaker of Serbo-Croatian from grammar books and from an orthographic dictionary. The verb forms were split into bases and endings in such a way as to reduce the number of bases to a minimum. The endings were then classified on the basis of their grammatical function, the bases were assigned to classes on the basis of their occurrence with the different classes of endings. Both base classes and ending classes were coded.

Summary of the Comparative Research on Adjectivals

The number of inflectional endings for adjectives in each of the four languages is:

Russian	39
Czech	43
Polish	27
Serbo-Croatian	18

The length of the inflectional endings ranges from zero to four graphs.

The establishment of this information naturally depends largely on the form of the transliteration system.

The base classes for adjectivals were established in terms of the distribution of the inflectional morphemes.

The number of classes established for each language was:

Russian	11 classes
Czech	9 classes
Polish	9 classes
Serbo-Croatian	7 classes

The closeness of the number of stem classes in the four languages indicates that there is a high degree of similarity in adjectival inflection.

Gender

All three genders are always distinguished by inflectional endings in the nominative and accusative singular in all four languages.

The distinction between the masculine and neuter on the one hand the feminine on the other is marked in the oblique cases in the singular in all four languages.

All three genders are distinguished in the nominative and the accusative plural in Serbo-Croatian, and also in Czech except in one paradigmatic class; Polish, all three genders are distinguished with the personal and impersonal aspect of the noun modified also constituting a basis for distinction.

Gender is not distinguished in the remaining cases of the plural in Serbo-Croatian, Polish or Czech.

Gender is not distinguished in any case of the plural in Russian.

Animateness

The category of animateness affects only the accusative case in the masculine singular or plural, and in the feminine plural. There are three possibilities. The accusative may have the same form as the nominative; it may have the same form as the genitive; it may have a form distinguished from either nominative or genitive. Two types of complication are introduced by special cases. Polish distinguishes personal and non-personal aspects in the masculine animate plural. Czech has a special adjective class (class AB) which has a different distribution from that of other Czech adjectivals in the masculine plural animate. This class also exhibits the peculiarity that the accusative has the same form as both the nominative and the genitive in the feminine and neuter singular; although this peculiarity has nothing to do with animateness, it is indicated here for the sake of interest.

The following table shows the form of the accusative in relation to animateness in the four languages. The symbols used are C: Czech in general; C': Czech adjective class AB; C": Czech adjective classes except AB; P: Polish in general; P': Polish personal; P": Polish impersonal; R: Russian; S: Serbo--Croatian.

Accusative in the		Like:	Nominative	Genitive	Neither	Both
masc.	sing.	inan. .	CPRS	-		
masc.	sing.	anim.	-	CPRS		
masc.	plur.	inan.	CPR	S -		
masc.	plur.	anim.	C'P"	P'R	C"S	
fern.	sing.		-	-	C"PRS	C'
fem.	plur.	inan.	CPRS	-	-	
fem.	plur.	anim.	CPS	R		
neut.	sing.		C"PRS	-	-	C
neut.	plur.		CPRS	-		

The Function of Inflectional Morphemes

The functions of an inflectional morpheme are determined by the distributional class of the stem with which it occurs.

Some inflectional morphemes are monovalent, and have only one function. Other inflectional morphemes are polyvalent, and have a combination of functions.

The total number of such functions and combinations of functions in the four Slavic languages under study is 121.

The number of functions which are equivalent in all four languages is 4 (3. 3 percent). The number of functions which are equivalent in three languages is 4 (3. 3 percent). The number of functions which are equivalent in two languages is 4 (3. 3 percent).

The number of functions which are different is 109 (ninety percent).

The high number of functions which are different is due to the fact that the adjectival inflectional morphemes are so often polyvalent. Nominal inflectional morphemes are more usually monovalent.

If the analysis of adjectivals were limited only to the case and number categories -- as it usually is -- then Russian exhibits only 13 patterns of polyvalency, as against 39 patterns when the categories of animateness and gender are also considered.

A similar pattern is found in the other three languages too.

This fact leads to the assumption that the number of grammatical categories under comparison directly affects the number of patterns of polyvalency in comparative work.

FRENCH-TO-ENGLISH MACHINE TRANSLATION RESEARCH

.

Nine tenths of the work of developing the French-to-English translation system was finished before June 1959. A brief history of that work will be given here.

In December 1956 and January 1957, 220 consecutive sentences from French chemical journal were studied; a vocabulary and a set of rules were built up on some two thousand filing cards. This vocabulary and these rules, if applied to the sentences mechanically by human beings, produced a fairly good English translation. It was a small beginning but this highly empirical method of attack seemed most promising. It appeared that the most productive second step would be to program the rules and vocabulary already in use, so that their insufficiencies could be easily and accurately exposed by testing them both on the original sentences and on further samples of text.

The only computer on which time was available free of cost for this experiment was the ILLIAC, at the University of Illinois. This machine has 1,000 forty-bit words of electrostatic storage in which the program active at any one moment is contained, and about 10, 000 words of magnetic drum storage. Since paper tape is the only medium of input and output, an automatic lookup in a large dictionary was not feasible. The lookup had to be simulated by hand, and the slowness of this procedure limited the rate at which sentences could be tested.

In retrospect, the 1, 000-word main memory was a blessing in disguise. A computer with a large program storage would have permitted programming the translation rules in straight-forward machine language. The 1, 000-word memory, however, made it necessary that the rules be compressed, even under the circumstances that a large part of the drum storage as well was used for the programs and these were copied into the main memory as needed.

Thus the rules, which had first been written in plain English, were rewritten in terms of elementary functions, in order to establish how large a set of such functions would be necessary. The elementary functions (which totalled about a hundred) and the corresponding sub-routines were easily contained in the main memory of the ILLIAC. Besides the sub-routines and a little work space, the main memory had also to contain an interpretive program that could locate macro-commands stored on the drum and execute them in proper sequence. This was the original form of the Simulated Linguistic Computer programming system (SLC).

By the fall of 1957, this very simple French-to-English algorithm had been programmed for the ILLIAC. Then some machine time on the IBM 704 at the National Bureau of Standards was offered to Georgetown. The mechanical and geographical advantages were so considerable that the program was rewritten for the 704, which had, at that time, 4000 words of core storage and no drum. Early in 1958 it was obvious that the simulation by hand of the

dictionary lookup was intolerably slow. Although a dictionary lookup program would actually be more difficult to write than the program for operating the linguistic algorithm, it seemed unavoidable to invest the necessary months in this work. By the fall of 1958, there was a 704 program which carried out all phases of an automatic translation program—reading a French text that had been keypunched in a natural way, and writing out the English equivalent.

A text of about 200, 000 words, taken from French journals of physics, was keypunched, and a crude concordance was made. A dictionary which included all the words in the text was coded, and new rules were based on such a study of the text as could be made without a complete concordance and within the available period of nine months. For various reasons, it was desirable that the system should be completed at whatever level of quality could be achieved by June 1959, and the system today is substantially the same as the system then. The dictionary contains about 4, 000 stems.

During the remainder of 1959, the fundamental programs of the SLC system were reprogrammed for the 709(0) computer; the existing French dictionary and translation algorithm did not need to be altered. In 1961, during the SLC coding of the Georgetown Russian-to-English translation system, a symbolic SLC coding language was devised and an assembly program was written for converting it into SLC numerical coding. A few months later, the interpretive program which is the core of the SLC was completely revised, and the symbolic assembly program was revised in parallel with it. This meant that the Russian-English material, being coded in symbolic form, needed only to be reassembled with the new version of the assembler.

The French-English material, however, antedated the symbolic system and existed only in numerical form. In order to avoid maintaining the earlier version of the SLC system, and in order to simplify any future work that might be done on French, some work was done on the French-English system. This consisted in rewriting all the independent operations and all the dictionary entries that contained local operations in the symbolic coding language that was in use for Russian. These operations were assembled into the new numerical language and inserted into the system, so that the same basic SLC programs now serve for both the French-to-English and the Russian-to-English systems. This work occupied several months in 1962. In 1962 also, a complete concordance of the French text was made as an aid to future work on the system.

The linguistic side of the French-to-English system is not at all easy to summarize. The work was begun with something of a bias against the application of academic linguistic analysis to machine translation. It seemed more promising to investigate completely one sentence at a time, in order to see what new rules each new sentence necessitated. If this approach is tempered with common sense, it exhibits points of superiority over the approach that begins by trying to formulate a complete grammar of French technical prose. A word-for-word translation of a French sentence into English, given certain easily formalized repairs, is usually acceptable. Instead of analyzing French, synthesizing English and then linking the two fundamental processes, it is

necessary simply to list the repairs. On the other hand, for translating Japanese into English by machine, this optimistic approach would be worthless.

General Operations

Some of these repairs in the French-English translation are of general application; whereas an adjective usually follows a noun in French, it usually precedes the noun in the English translation. Some of the repairs are particular to individual words; 'chemin de fer', for examples, is to be translated 'railroad', not 'road of iron' nor 'iron road'. The more general repairs in the system, in the order in which they are carried out on a sentence, are listed here: Words which are ambiguous as to part of speech have this ambiguity - resolved. The patterns of such ambiguity and some examples are: ,

	Type	Examples
a)	article/pronoun	le, la, 1', les,
b)	preposition/pronoun	en
c)	verb/preposition	entre
d)	verb/adjective	présente
e)	verb/noun	porte
f)	3rd person present tense verb/past participle	réduit
g)	types (e) and (f) combined	fait, produit

Some additional types are held for later discussion.

First the program resolves all ambiguities that can be cleared up by examination of the immediate context; this includes practically all instances. One variety of exception occurs when an ambiguity of type (a) is followed by an ambiguity of type (d) or (e); the combination is ambiguous since it can be read either as article plus noun or as pronoun plus verb. For example, in the combination 'le produit', the two words agree in gender and number if they are read as article and noun, and there is no pronoun or preposition immediately preceding which might resolve the ambiguity. The sentence traditionally quoted as an example of this sort of ambiguity is 'Le pilote ferme la porte'. This can mean either:

(article-noun): 'The pilot shuts the door', or
(pronoun-verb): 'The strong pilot carries it. '

Actually, such instances of unresolvable ambiguity at the syntactic level are uncommon. The resolvable instances are settled by choosing the article-noun combination unless something to the left in the sentence can serve as the subject of the verb in the pronoun-verb combination. This blanket rule works well, but the formulation of a better rule is undoubtedly

possible. A routine that analyzes every sentence so as to mark the clause boundaries, the subject blocks and the predicate blocks as efficiently as possible would be well worth developing, not only for itself, but because of other characteristics of the system. At present, the subject and predicate may be identified by various routines, including the ambiguity-resolution routine described above. When such an identification is made, it is not necessarily coded. Consequently, another routine may repeat the work of making that same identification. Or again, the identification of a combination of an auxiliary verb with a past participle may be repeated many times in this translation system. Yet it is easily possible to make these identifications at an early stage, and to code them so that the information will continue to be available during the whole translation process.

(This digression has been made to show how theories on method of attack have been modified since 1957. It is now recognized that certain kinds of grammatical analysis can be profitably applied to every sentence. On the other hand, an oversystematic first approach might have bogged down in the analysis of difficult sentences, many of which can be translated into English on a word-for-word basis.)

2. Whenever an adverb intervenes between a preposition and an infinitive governed by the preposition, the adverb is moved to a position after the infinitive. If an infinitive is preceded by 'a', and if the translation of the preposition has not been fixed for some other reason, the translation of 'a' is altered from 'at' to 'to'. If an infinitive is preceded by 'de', and if the translation of the preposition has not already been fixed, and if the preposition is not preceded by a noun or adjective, the translation of 'de' is altered from 'of to 'to'. After these changes have been made, if any infinitive is preceded by a preposition whose translation is not 'to', the infinitive is marked so that it will continue to function grammatically as an infinitive but will finally be translated as an English gerund. Thus, 'a time to wash' has 'wash' because of 'to', but 'a time for washing' has washing because of 'for'. Of course, 'to' may also occur with the gerund as in 'It amounts to washing in mud'. In such a case it is possible to deceive the program by inserting a dummy character in the coding of 'to'. The dummy character induces the use of the gerund and then is dropped when the translation is written out.

3. The necessary inflection of English verbs and nouns is carried out. Any problem of choosing among different possible English equivalents will almost always have been solved by the time this point is reached. The French and English verb tense systems are sufficiently similar to make the choice of the English tense form easy. It is a fortunate characteristic of scientific reports that no outrage is committed if the French imperfect is always rendered by the English simple past tense. Similarly, the French perfect tense, which is often used in French scientific articles where a more literary style would call for the simple preterite, can safely be rendered by the English perfect tense. Negation is handled as a special kind of verb inflection.

4. Whenever a singular noun is immediately preceded, but not immediately followed, by some punctuation other than quotation marks, the English indefinite article is inserted before the noun in such a way that adjectives can still be inserted between this article and the noun.

5. Whenever a present participle is immediately followed by an article, the participle is marked so that it will not be treated as an adjective and will not be shifted to a position after the article. Thus, 'the wheel turning the screw' will not be changed to 'the wheel the turning screw'. This rule suggests several more difficulties of order connected with present participles; these difficulties are ignored in the system but fortunately they do not seem to be exemplified very often in scientific articles.

6. French adjectives can be divided into a small class that precede the nouns they modify, and a much larger remainder that follow. But some of each class are inconsistent, and an attempt is made at this point in the program to alter the grammatical codes of adjectives if necessary. Thus: while 'éternel' usually follows its nouns, it precedes in '1'éternel retour'; the 'new franc' has been called both 'le franc nouveau' and 'le nouveau franc'.

7. Adjectives modifying a noun they follow, along with adverbs modifying those adjectives, are shifted to a position preceding the noun. This is straightforward except in cases like 'les officiers et les soldats malades' and 'un fil de fer blanc'. The former difficulty is fairly common and seems intractable for the moment. In the case of two nouns joined by 'de' and followed by an adjective that could agree with either, the rule adopted in this: if the second noun is preceded by an article, the adjective will be taken as modifying it; otherwise the adjective will be taken as modifying the first noun. Of course, both 'fil de fer' : wire, and 'fer blanc' : tin, have to be handled by individual rules as "idioms". If the phrase 'un fil de fer blanc' occurred, no doubt the two idiom rules would conflict and spoil the translation.

8. Whenever a verb in the first person plural of the present tense stands first in a sentence, or immediately follows a mark of major punctuation, the words 'let us' are prefixed to its translation.

9. Object pronouns are moved from their position before the verb that governs them to a position after the verb, or after the past participle with which an auxiliary verb is connected.

10. The word 'to' is inserted before any infinitive not preceded by a preposition.

Local Operations

The foregoing may seem like a very sketchy set of rules for converting a word-for-word translation of a French sentence into a more acceptable translation. However, the bulk of the work is done, not by rules applied to every sentence, but by rules associated with particular words. The rules in a local operation are to be applied in the order given, and this is to be understood from the word 'otherwise' which precedes all rules but the first.

There is a local operation connected with the word 'bien', which, like most of the local operations, is applied immediately after rule 1 of the general operations. The rule is approximately as follows:

'bien'

1. The collocation 'si bien que' is to be replaced by a single theoretical conjunction; translation: 'so that'.

2. Otherwise, 'si bien' will be read as two discrete adverbs; translation: 'so well'.

3. Otherwise, 'ou bien' will result in the deletion of 'bien', and the alteration of the translation of 'ou' from 'or' to 'or else'; translation: 'or else'.

4. Otherwise, 'aussi bien. . .que', if no punctuation or verb intervenes between 'bien' and 'que', will result in the deletion of 'aussi bien' and the replacement of 'que' by a theoretical conjunction which does not have the rules associated with 'que'; translation: 'as well as'.

5. Otherwise, 'très bien' will be transferred word for word; translation: 'very well'.

6. Otherwise, 'assez bien' will cause the translation of 'assez' to be altered to 'fairly'; translation: 'fairly well'.

7. Otherwise, if 'bien' is preceded by a preposition, translation: 'well'.

8. Otherwise, if 'bien' is immediately followed by a past participle:

    a) if there is a form of 'avoir', or if there is a form of 'être' preceded by a reflexive pronoun, to the left of 'bien' and only adverbs intervene, translation: 'indeed'.

    b) Otherwise, if the past participle is not 'entendu', translation: 'well'.

    c) Otherwise, if the past participle is 'entendu', and if it is not followed by 'que', replace 'bien entendu' by a single theoretical adverb; translation: 'of course'.

d) Otherwise, 'bien entendu que' must occur; change the translation of 'entendu' to 'understood'; translation: 'well'.

9. Otherwise, if 'bien' is immediately followed by 'sûr':

a) if 'sûr' is immediately followed by 'que', translation: 'quite'.

b) Otherwise, 'bien sûr' is to be replaced by a single theoretical adverb; translation: 'of course'.

10. Otherwise, if 'bien' is immediately followed by 'certain', translation: 'quite'.

11. Otherwise, if 'bien' is immediately followed by 'plus'; 'moins', 'meilleur', 'pire', 'mieux', 'pis', or 'd'avantage', translation: 'much'.

12. Otherwise, if 'bien' is immediately followed by an adjective, translation: 'very'.

13. Otherwise, if 'bien' is immediately followed by '-fondé', 'bien-fondé' is to be replaced by a single noun; translation: 'soundness'.

14. Otherwise, if 'bien' is immediately followed by 'que', 'bien' is to be deleted and the translation of 'que' is to become 'although'.

15. Otherwise, if 'bien' is immediately preceded by a verb, translation: 'indeed'.

16. Otherwise, translation: 'well'.

(It may be of interest that this rule is entirely contained in 85 words of computer memory.)

Intermediate Operations

Intermediate in generality between the local operation cited for 'bien' and the general operation applied to every sentence, there are rules associated with sizable numbers of words, and these rules may leave 'parameters' to be set in individual cases. For instance, there is an operation associated with many verbs that runs:

> If this verb is followed in the same clause by the preposition X, and if no direct object intervenes between the verb and the preposition, alter the translation of the preposition to Y (unless it has already been fixed by an earlier rule) and alter the translation of the verb to Z.

In the dictionary entry for each of these verbs, an instruction calling on this will be coded, and it will be followed immediately by the appropriate values for X, Y and Z.

A few other intermediate operations are connected with ambiguities as to part of speech. Ambiguity types 1 to 7 were listed under general operations as operation 1. Ambiguity types 8 to 13 are handled by intermediate operations. They are:

8. masculine adjective/noun : minéral
9. feminine adjective/noun : caracté'ristique
10. masculine present participle/noun : composant
11. feminine present participle/noun : gouvernante
12. masculine past participle/noun : composé
13. feminine past participle/noun : partie

Type 12, for example, is handled by an intermediate operation, which is associated with verb stems like 'compose-':

> If the word containing this rule is a past participle masculine, and if the word stands first in the sentence, or if it is immediately preceded by an article, a preposition, or an adjective of the type that normally precedes the noun it modifies, change the grammatical codes of this word from past participle to noun, preserving its gender and number, and change its English equivalent to X.

ENGLISH TO TURKISH TRANSLATION RESEARCH

The English-to-Turkish translation research was undertaken with two aims in view. One aim was, of course, the achievement of English-to-Turkish machine translation. The second aim was the development of so broad an analysis of English that the system might also prove useful in the synthesis of English for Russian-to-English translation and in the analysis of English text for information retrieval.

Because of the above aims, the English analysis was made as complete and as independent as possible. Only when the analysis of English as English is complete is the process of synthesizing Turkish considered. This dichotomy in no way detracts from the efficiency of the English-to-Turkish translation system; indeed, it provides from the beginning a great deal of information from the English which would undoubtedly be found necessary in the long run for any refined Turkish translation, moreover the dichotomy readily allows of the application of the English analysis to the translation of English into other languages, as well as to the other uses suggested above.

The translation system can be divided into two parts; the dictionary and the operations.

The Dictionary: The dictionary format sheets are arranged so as to permit of entering information under four broad categories. These are: entry, entry coding, gloss and gloss coding. The coding is subdivided under the headings morphological, syntagmatic and semantic coding. Each entry may have any number of codings; each coding may have any number of glosses; each gloss may have any number of codings.

Entry; All entries are full entries at present. There is every intention of developing a split entry system at a later date. The present dictionary, however, holds some seven hundred formally distinctive items. The introduction of a split entry system in such a small dictionary, with the consequent necessity of developing and programming a morphology routine to match the stems with the inflectional endings, will involve considerably more work than the simple process of listing every formally distinctive item separately as a full form.

Entry Coding: The morphological level of coding is not in use as yet, since all entries are full entries.

The syntagmatic level of coding is at present the most highly developed level. All entries are classified as parts of speech, and eight parts of speech are distinguished. These parts of speech are designated by the digits from 0 through 7. It has on occasion seemed advisable to establish other parts of speech by promoting certain classes of words which are now listed as subcategories of established parts of speech. This has not yet been done, however, and the system will be described as having the following eight parts of speech.

0. Punctuation: The marks of punctuation are the full stop (1), the question mark (2), the exclamation point (3), the semi-colon (4), the colon (5), the comma (6), the dash (7), the parenthesis (8), and the quotation marks (9). Of these, one through five are marks of final punctuation. Six (the comma) is ambiguous. Seven through nine are parenthetical; parenthetical punctuation marks are further sub-classified as opening a parenthesis (1) or closing a parenthesis (2) . Special routines are necessary to decide whether any comma is to be treated as a mark of final punctuation, as a mark of parenthetical punctuation, or as fulfilling some other, generally non-significant, function.

1. Nouns: Nouns are at present divided into six classes, but it seems reasonable to assume that further sub-classification will be necessary. These classes may be described by rubric as countable, uncountable, specific, count, plurale tantum, and eponymous. Nouns are also cross-classified into singular but not possessive, plural but not possessive, singular and possessive, and plural and possessive. Nouns are cross-classified again into neuter, masculine, feminine and common. It seems that a further cross-classification of nouns into non-personal and personal will be eventually useful, but the text on which this research is based does not require it.

2. Verbs: Verbs are divided into four classes in accordance with certain features of order which become apparent when the verbs occur together in the same structure. Verbs are cross-classified into six classes on the basis of form; these classes are best exemplified by the forms of the verb 'sing': sings (1), sing (2), sang (3), to sing (4), singing (5), and sung (6). Certain sub-classifications are made in the case of the verb 'to be' where the forms 'am' and 'are' equate with 'sing', and where the forms 'was' and 'were' equate with 'sang'. A certain few verbs have identity codes so that they can be individually recognized if necessary.

3. Adjectives: Adjectives have been provisionally divided into seven classes on the basis of the order in which they occur when more than one of them modify the same noun. Some of these classes are further subdivided to indicate various refinements of the order characteristics. All adjectives are cross-classified in accordance with the manner in which they form their comparative and superlative.

4. Adverbs: Adverbs have been provisionally divided into seven categories in terms of the order in which they occur when more than one of them modify the same verb, as well as in terms of whether they modify adjectives or other adverbs, and in terms of their possible relationships to the syntactic units (subject, verb, sequence). Adverbs are cross-classified in terms of the manner in which they form their comparative and superlative. A few adverbs, such as 'not', have an identity code so that they may be individually recognized if necessary.

5. Prepositions: Prepositions have not as yet been classified. Each preposition, however, has an identity code so that it can be individually recognized if necessary.

6. Conjunctions: Conjunctions are subdivided into four groups: coordinating conjunctions, subordinating and participating conjunctions, subordinating and non-participating conjunctions, and disjunctions. In addition, each conjunction has an identity code so that it can be individually recognized if necesssary.

7. Pronouns and Surrogates: Pronouns are subdivided into three classes: first person, second person, and third person. The are cross-classified, in terms of whether they are singular, plural, or amphoteric. They are also cross-classified in terms of whether they are neuter, masculine, feminine, or common. They are further cross-classified in terms of form: first form (I), second form (me), third form (mine), and fourth form (myself).

The whole question of noun and verb surrogates remains open. Their classification will undoubtedly be necessary in a complete English analysis, but the text has so far not required it.

In addition to the coding for parts of speech, each word can be coded to show its sequence; its sequence is the sum of the limitations on the type of dependencies which occur in a structure of which this particular word is the head. Thus, each verb is coded to show whether it requires no object, one object, two objects, a complement, or an object with a complement. Certain subcodings are introduced to indicate specific restrictions in the nature of either object or complement. Each verb is also coded to show what prepositions introduce dependent prepositional phrases, and whether these prepositional phrases are bound or facultative. Similar coding is possible for any other part of speech, though there are fewer of the other parts of speech which are so idiosyncratic in their sequence requirements.

The semantic level of coding has not as yet been developed, although plans are being made for developing codings for restricted areas of the vocabulary which associate with certain structural classes, especially with prepositions; semantic codings are also projected for a limited type of field-of-discourse classification.

Gloss: The Turkish gloss is right justified, and is entered as a split form if it is subject to inflection. The Turkish inflectional suffixes are also entered as glosses. In some cases, where the entry is inflected, the gloss consists of both stem and inflectional ending; this is entered as one gloss, but stem and ending are put in separate areas so as to make possible the addition of other endings with a lower order number than that of the ending which forms part of the gloss. In other cases, where the entry is an adverb and the gloss is a case form of a noun, stem and inflection may be entered as a unit, since no adding of other inflections is to be expected.

Gloss coding: The morphological level of coding is very highly developed. All entries are coded to indicate whether they are stems or suffixes. If they are stems, they are coded to indicate whether they are noun, verb, adjective or invariable. Invariables are subdivided into adverbs and prepositions. The question as to whether pronouns are to be recognized as distinct from nouns

and whether numbers are to be recognized as distinct from adjectives is still open. (The four Turkish students who attend the seminar on machine translation are currently preparing papers on subjects which will deal directly with this question.) Every stem or suffix is coded to show what vowel harmony it requires, if indeed it requires any. All suffixes are coded to show what allomorphic variations may occur as a result of combination with certain types of stems. In the same way, all stems are coded to show whether they accept the regular pattern of suffixes, or some specific variation. All stems and suffixes are coded to show what their finals are, and how these finals vary allomorphically when suffixes of certain types occur with them. The syntagmatic level of coding shows in a rudimentary fashion what syntagmatic sequences occur with that stem. (It is expected that some of the papers to be produced from the machine translation seminar will be useful in developing more extensive sequence coding.) The semantic level of coding has not been developed at all, and there is some question as to whether it will be useful or necessary as long as Turkish remains a target language.

The Operations: The operations of the English-to-Turkish translation system are divided into two categories. These are the English analysis and the Turkish synthesis.

The English Analysis: When the dictionary lookup has been completed, the present linguistic statement calls for the performance of five major operations. There are also some minor operations which can be called upon at any time during the process of analysis. Those which are called upon most frequently during some particular major operation are mentioned in connection with that major operation.

The major operations, with the significant minor operations, are segmentation (and correlation), preposition blocking (and infinitive blocking), verb blocking, subject blocking, and sequence blocking.

Segmentation: The process of segmentation divides the text, where necessary, into machine sentences. A machine sentence may be defined as one which contains a single verb structure with its subject and sequence.

The segmentation operation begins at the beginning of the text, or at the last mark of final punctuation which has been recognized as a segmentation boundary. It proceeds to the right through the text, counting and recording commas and establishing the function of the parenthetical marks of punctuation, until a mark of final punctuation is reached. This mark of final punctuation is marked as a segmentation boundary. The routine then moves to the left through the segment which has just been delimited and inspects each comma in terms of its relationship to conjunctions and participles, establishing other segmentation boundaries where certain specified combinations of factors occur. The routine then moves to the right again and inspects each conjunction which is not preceded by a comma, and establishes further segmentation boundaries as necessary.

The subordinate routines which are most frequently called into play during the segmentation routine are the correlation routine and certain of the ambiguity routines. The correlation routine decides the status of correlative conjunctions as conjoining sentence structures, partial sentence structures, or non-sentence structures. The ambiguity routines decide such questions as whether words like 'for' or 'since' are conjunctions or prepositions in any particular case.

Preposition Blocking: Preposition blocking searches out the prepositions in each segment, resolves any ambiguity that occurs, and calls upon the Noun Structure from the Left Operation to establish the noun structure governed by the preposition. The maximum possible noun structure is always established, but in certain cases where other analyses are possible, special indications are given so that these noun structures may be reconsidered at a later stage if the maximal analysis interferes with verb or object blocking. Because of the ambiguity of 'to' as either preposition or infinitive marker, infinitive blocking also, is most conveniently performed at this stage in the operation.

Verb Blocking: Verb blocking is carried out before either subject or sequence blocking for two reasons. First, the verb block is usually easy to recognize. Second, verb blocking usually divides the sentence into three sections, of which the verb block is the center section; the first section is statistically almost certain to contain the subject, and the third section is statistically almost certain to contain the sequence.

Verb blocking is accomplished by searching first for class I verbs, and then for class II, III, and IV verbs in that order. If there are no class I, II, or III verbs in the segment, there may be more than one candidate for a class IV verb, and the routine provides mechanisms for considering the qualifications of each candidate, and for selecting the most appropriate.

Subject Blocking: At this stage of the operation, the subject is usually easily recognizable, and there remains only the problem of delimiting the subject block. This is achieved by calling on the minor operation: Noun Constructs from the Right.

Sequence Blocking: The syntagmatic coding of each verb indicates whether objects or complements are to be expected, and the number of objects or complements which may occur. If a number of homonymous verbs have different sequence codes, the test is made first for the verb with the longest possible sequence, and then for the verbs with increasingly shorter sequences until a verb whose sequence coding fits the particular sequence is found. Possible ambiguity arises in the case where a verb sequence contains an object or complement followed by a prepositional phrase. No sure and general method has been evolved as yet for determining whether this prepositional phrase qualifies the nominal structure which it follows, or modifies the verb block of the segment. In some cases, the syntagmatic coding of the verb proves useful; in other cases, semantic coding will undoubtedly prove useful, but has not yet been developed; the present solution for the difficulty is to establish a blanket rule that the prepositional phrase will be construed as qualifying the nominal construction unless

the verb sequence coding specifically indicates that it modifies the verb. This blanket rule is about 66% effective. There is no doubt but that further research will lead to a more generalized statement which will be considerably more effective.

Segment Correlation: In text sentences which are composed of a number of machine sentences which are subordinate clauses, the structural function of the subordinate clauses in the main sentence must be established and coded. Sometimes this has been done in the course of one of the preceding operations. If it has not, it is done here.

At this point the English analysis is complete.

The Turkish Synthesis: The Turkish synthesis is divided into five operations at present. These have been named the Annexation Operation, the Syntagmatic Rearrangement Operation, the Syntactic Rearrangement Operation, the Transfer Operation, and the Self-Organization Operation. The order of the first four operations is immaterial; the first three operations are really subdivisions of an overall rearrangement operation and could probably be fused into one.

The Annexation Operation: The Annexation Operation recognizes noun-noun constructs in English, such as 'wire loops' or 'current flow', and also noun-'of'-noun constructs, such as 'loops of wire' or 'flow of current'. The items occurring in these constructs are internally rearranged when necessary and have indications added to them of the Turkish suffixes which will be necessary, depending on the makeup of the individual construction, and of the adjectival dependencies of each noun.

The Syntagmatic Rearrangement Operation: The Syntagmatic Rearrangement Operation is intended to cover all problems of rearrangement which are not concerned with annexations, or with block rearrangement. Actually, it is chiefly the English prepositions that are rearranged at this point. Their position is changed so that they come to follow their object.

The Syntactic Rearrangement Operation: The Syntactic Rearrangement Operation rearranges the blocks of each segment into the following standard order: subject block, adverb block, object block, and verb block. Where individual constructions in the adverb block are marked on the Turkish side as being of such a class, the adverbial construction is assigned to a position before the subject block. Research being done at present indicates that it may also be necessary to establish a third class of Turkish items which translate English adverbs and which will be assigned to a position between the object block and the verb block.

The Transfer Operation: The Transfer Operation focuses exclusively on the Turkish glosses, and resolves any outstanding problems in the choice of gloss. Those glosses which are Turkish suffixes and which occur in groups are re-arranged in the order in which they are to occur. This is done by means of the suffix order codes against each such gloss in the dictionary. All suffixes

are given in a canonical form which shows their potentialities for allomorphic variation.

The Self-Organization Operation: The Self-Organization Operation binds each Turkish stem and its endings into one word; the rules of vowel harmony and of allomorphic variation are applied; the result is printed out as the Turkish translation.

In order to show the progress of a sentence through the various operations of English analysis and Turkish synthesis, first a simple sentence and then a complex sentence from the original text are given at the end of this paper.

The research which has led to the translation system briefly outlined here uses as its basis an actual English text in the field of electricity and magnetism; a man-made translation into Turkish of the English text serves as a guide to the level of translation to be attempted.

The entire English text consists of about 3500 words. A segment of the text, some five hundred words in length, was chosen as the basis for a preliminary study. After the problems presented by this segment had been explored, the embryonic system was expanded so that it could cope with the problems presented by the entire text.

This report represents a point in the development of the research where most of the basic problems presented by the entire text have been covered.

Because of the difficulties involved in communications between Washington, where most of the research is being carried out, and Ankara, where the guide translation was prepared and where the dictionary is being coded, and also because of lack of funds for this specific research, it has not yet proved possible to program this translation system and to test it on a machine. It is hoped that in the course of the next year it will be possible to find the necessary funds for testing the system and for adjusting it or reworking it until it can produce usable translations.

Example I

Text:

The thumb will then point to the north pole of the coil.

Segmentation:

@@ The thumb will then point to the north pole of the coil. @@
(There is no internal segmentation in this sentence.)

Preposition Blocking:

@@ The thumb will then point to the north pole of the coil. @@

p/ a a n p/ a n
P P

Verb Blocking:

@@ The thumb will then point to the north pole of the coil. @@
vl/ d/ v4 P P
V D V

Subject Blocking:

@@ The thumb will then point to the north pole of the coil. @@
a n V D V P P
S

Sequence Blocking:

@@ The thumb will then point to the north pole of the coil. @@
S V D V D

Annexation Operation:

@@ The thumb will then point of the coil to the north pole -X. @@
S V D1 V D2

Syntagmatic Operation:

@@ Thumb then point-will coil-of north pole -X -to. @@
S D1 V D2

Syntactic Operation:

@@ Then thumb coil -of north pole -X -to point -will. @@
1 2 3 4 5 6 7 8 9 10

Transfer:

@@ BU TAKDIRDE BASQPARMAK BOBIN -4N KUZEY
1 2 3 4 5

KUTUB -4 -N4 GOWSTER -2C2K -D4R. @@
6 7 8 9 10

Self-Organization:

@@ BU TAKDIRDE BASQPARMAK BOBININ KUZEY KUTBUNU GOWSTERECEKTIR. @@

Example II

Text:

The internal lines are concentrated within the area of the core, which becomes strongly magnetic, having a north and south pole the same as a permanent magnet.

Segmentation:

@@ The internal lines are concentrated within the area of the core,
@(which becomes strongly magnetic)@
@/ having a north&and&south pole the same as a permanent magnet /@@

(The text sentence has been subdivided into three machine sentences. The first machine sentence is the main clause, and may be considered to include the second and third machine sentences as dependencies. The second machine sentence is a subordinate clause having an adjectival function (this fact is marked by the parentheses). The third machine sentence is a participial construct (this fact is marked by the virgules) and is ambiguous as to function, but it will finally be resolved as having an adjectival function also. The items 'north' and 'south' have been correlated as a non-sentence construct.)

Preposition Blocking:

@@ The internal lines are concentrated within the area of the core,
p /a n — P; p /a n — P

@(which becomes strongly magnetic)@
@/ having a north&and&south pole the same as a permanent magnet. / @@
p /a a n — P

Verb Blocking:

@@ The internal lines are concentrated within the area of the core,
v3 v4 — V; within the area — P; of the core — P

@(which becomes strongly magnetic) @
v4 — V

@/having a north&and&south pole the same as a permenent magnet. /@@
v45 P
VP

Subject Blocking:

@@The internal lines are concentrated within the area of the core,
a a n V P P
S

@(which becomes strongly magnetic)@
c2 V
SR

@/ having a north&and&south pole the same as a permanent magnet. /@@
VP P

Sequence Blocking:

@@ The internal lines are concentrated within the area of the core,
S V D

@(which becomes strongly magnetic)@
SR V d a
C

@/ having a north&and&south pole the same as a permanent magnet. /@@
VP a a n P
O D

Segment Correlation:

@@ The internal lines are concentrated within the area of the core,
S V D

@(which becomes strongly magnetic) @
SR V C
A (core)

@/(having a north&and&south pole the same as a permanent magnet.)/@@
VP O D
A (core)

Annexation Operation:

@@ The internal lines are concentrated of the core within the area-X,
S V D

@(which becomes strongly magnetic)@
SR V C
A (core)

@/(having a north&and&south pole the same as a permanent magnet.)/@@
VP O D
A (core)

Syntagmatic Rearrangement:

@@ Internal lines concentrated are core -of area -X -within
S D V

©(which becomes strongly magnetic)@
SR V C
A (core)

@ /(having north& and& south a pole same permanent a magnet -as)/@@
V P O D
A (core)

Syntactic Rearrangement:

@@ Internal lines @/(same permanent a magnet -as north& and& south a pole
1 2 3 4 5 6 7 8 9 10 11 12

having) / @(strongly magnetic becomes -which)@ core -of area -X -within
13 14 15 16 17 18 19 20 21 22

concentrated -are @@
23 24

Transfer:

ICQ HAT -L2R TXPKX DAIMI BIR MIKNATXS GIBI KUZEY VE GUWNEY
1 2 3 4 5 6 7 8 9 10

KUTUB -L2R -2 MALIK OLARAK KUVVETLE MANYETIK HALE GEL -EN
11 12 13 13 14 15 16 17

CQUBUK -4N SAHA -S4 DAHILINDE TOPLA -N -4R - L2R
18 19 20 21 22 23 24

Self-Organization:

ICQ HATLAR TXPKX DAIMI BIR MIKNATXS GIBI KUZEY VE GUWNEY KATUPLARA MALIK OLARAK KUVVETLE MANYETIK HALE GELEN CQUBUGQUN SAHASX DAHILINDE TOPLANXRLAR.

RESEARCH IN CHINESE

Mechanical translation from Chinese into a European language presents greater difficulties than translation between two European languages; it is hardly possible to list here all the characteristics that make Chinese so difficult to translate by non-intuitive procedures. But a striking difficulty is that, even before any problem of translation algorithms arises, the nature of the Chinese writing system enormously complicates any efforts to gather and collate information using data-processing machines. There is no known practical way of automatically deriving a digital code from a pictographic Chinese character, and therefore, at present, there is no way to introduce a Chinese text into a machine without the intervention of a person who knows Chinese.

The best way to put a Chinese text onto punched cards seems to be to use the telegraphic code. Every character in the language can be represented, for telegraphic and other purposes, by a four-digit code, the combinations ranging from 0001 to 9999. The Georgetown Machine Translation Research Project was fortunately able to find people who had had long practice with the telegraphic system and could encode rapidly.

Preparation of the Text

The first 115 pages of Mua Lo-Keng: An Introduction to Number Theory; Science Publishing Company, (no city given), 1957 were copied onto foolscap, one column to each page. Anything not a Chinese character nor an ordinary punctuation mark was represented by a cover symbol X, both because the mathematical substance of the book was of no direct interest, and because the question of how mathematical items differ in the way they affect the syntax of the sentences containing them must wait upon the solutions to many other problems before it can be answered. This text consists of about 20, 000 running Chinese characters, not counting the symbol X nor any punctuation mark.

Each character was then matched by its telegraphic code number in a parallel column on the foolscap copy, and the now encoded text was punched on cards. Concurrently, an English translation was made from the original Chinese text, with all the mathematical material in its original place.

Using these materials, two persons who are familiar with the Chinese language but unfamiliar with mathematics were able to mark character groupings on the foolscap copy of the text, and to write in approximate translations of the Chinese nouns and verbs in the text. Most of the function-words remained untranslatable, however. Each Chinese character group (whether a unit character or a phrase) was entered on a filing card, along with the equivalents assignable to them on the sole basis of the English translation. Some received one translation, some a number of alternative translations, and some simply a reminder of the pronunciation and grammatical function of the Chinese word.

When about one quarter of the text had been processed in the manner described, and a little more than five hundred character groupings had been collected, new character groupings seemed to be appearing rather slowly. And so, at this rather late stage, data-processing machinery was used. The character groupings already collected were keypunched, and a computer program
converted the cards into an ordered dictionary on magnetic tape. Another program read the entire text, looked up each character grouping in the dictionary as far as possible, and produced an output of which the following is a sample sentence:

2399	THEREFORE
X --------	
0035 -------	
0001	ONE
2419	WHOLE
0190 2422	COEFFICIENT
0037	(JR1)
1122 7309 1709	POLYNOMIAL
* ____	

This sentence was originally keypunched thus:
2399 X 0035 0001 2419 0190 2422 0037 1122 7309 1709 *

(For convenience in scanning listings by eye, the asterisk is used in the place of a period.) There is no point as yet in putting punctuation marks or X into the dictionary, so they are treated as characters missing from the dictionary, and receive six dashes to the right in the output. In the sentence shown, the only Chinese character missing from the dictionary was 0035, which was consequently also marked with six dashes on the right. The other character groupings were all looked up, either as individual characters (2399, 0001, 2419 0037) or in phrases (0190 2422 and 1122 7309 1709).

Various strategies might be chosen for grouping characters into phrases where possible; the following strategy appears to be the simplest and to work well, at least in the mathematical text:

1. Each sentence is treated independently of its neighbor.

2. The whole sentence is matched against the list of character groupings contained in the dictionary, and the longest grouping which matches a corresponding number of characters at the beginning of the sentence is accepted. For example, suppose a sentence begins with 0278 6677 1840 5287 . . . . . and suppose the dictionary contains the three following entries:

0278 6677 1840	TRANSITIVITY	(1)
0278 6677	TRANSITIVE	(2)
0278	TRANSMIT	(3)

Entry (1) is the longest matching dictionary entry, and is accepted. If entries (2) and (3) were in the dictionary but entry (1) were not, then entry (2) would be accepted; similarly, if entry (3) were in the dictionary but entries (1) and (2) were not, then entry (3) would be accepted. If none of the three is in the dictionary, the look-up program behaves as if there were the following entry in the dictionary for the first character:

0278----------

and accepts that entry.

3. The entry chosen is copied into the output.

4. If the sentence has been exhausted, the next sentence is treated in the same way.

5. If the sentence has not been exhausted, the program returns to step 2, and proceeds as if the sentence began with the character following the last character matched, whether that last character was matched as the last member of a phrase or as an individual character.

The potential weakness of this procedure will be seen if it is supposed that the dictionary contains not only the three entries shown above, but also entries for the phrase 1840 5278 and for the character 5278. The series of characters: 0278 6677 1840 5278.....could then be looked up either as 0278-6677-1840 5278 .....or as 0278-6677 1840-5278. The present program would invariably give the former result, but one can well imagine that the latter result would in many cases be the proper one. In the present mathematical text, however, with its restricted vocabulary, the difficulty does not seem to arise. With a larger and more general vocabulary a more complicated strategy might have to be adopted.

The output of the sample sentence reproduced above shows that character 0035 is present in the text but missing in the dictionary. The corresponding sentence in the English translation is "Hence GK(x) is a polynomial with integral coefficients." From this one can conclude that it will be useful for the time being to give the English equivalent 'is' to character 0035. A similar process is applied for the other characters and phrases missing in the dictionary. The first quarter of the text was processed slowly by hand and with constant consulting of the card file, to see whether any particular character grouping was already recorded; the remaining three-quarters of the text can now be processed by machine, and it is possible to collect the remainder of the translatable vocabulary quickly by comparing the machine output with the human translations.

Actually, the character 0035 is not a good example of this vocabulary-gathering procedure. It is not a very rare equivalent for 'is' or 'are'; the reason it is not already in the dictionary is that there seem to be occurrences that cannot be translated by the verb 'to be'. To make a further study of such problems, a concordance would be extremely helpful, and it is hoped that when the easier vocabulary for this text has been listed, the output from the present look-up program will be a useful foundation for making a concordance of the text.

Concordances

No work directly connected with concordance-making in a Chinese text has been done so far. The output from the present dictionary-lookup program cannot be described as even a "very rough translation", and any vocabulary listing completed rapidly with its help would admittedly contain only the easy words and phrases, along with indications concerning the grammatical particles that would be helpful only to someone acquainted with the language. For these reasons, the whole procedure may appear pointless. However, it is precisely to obtain concordances that the work undertaken so far has been done, and concordances should be of great value in coming to grips with the more difficult aspects of Chinese.

Exactly what is meant by 'concordance' here?

Here is a sample from a concordance of a French text:

```
que 1'étendue des $sym *u ou  il y a ressemblance entre les deux fi  27504
(u, v) a besoin d'être precise    il y a toujours coincidence suffisant  27195
ble du système de chocs, mais il y a toujours réversibilité micro-p 17746
e quantité de chaleur sans qu'  il y ait variation de température. $p  28518
de , . . . un peu avant la fusion, il y aurait formation de petits "noya  30333
```

For every word in the text, a line is created with the first letter of that word in the center of the line, as much of its environment reproduced to the right and left as the line allows. (Actually, more environment can conveniently be given in computer printout than is shown above, because printout paper is wider.) A reference to the position of the word in the text is given on one side. These lines are then sorted alphabetically, the first character of the key word in the line being the most significant for the sorting, and successive characters to the right being successively less significant. This is effectively the same as saying that the sorting is done by words, the keyword being the most significant, and successive words to the right being successively less significant. (A different sorting system could also be programmed, in which the key word is the most significant, and successive words to the left are successively less significant.)

In the sample of a French concordance reproduced above, the key word is 'il'. It is seen that the sorting system has grouped together all the instances of 'il y a' in the text, and indeed all instances of 'il y' followed by any other word. For studying the use of 'il y (avoir)' in the text, the value of this concordance is obvious. The value of something similar for studying Chinese texts would be even greater. Definitive linguistic descriptions of spoken Chinese dialects remain to be written, and a systematic description of the written language is not to be found. There seem to be many particles in written Chinese which a writer of the language can use correctly, but whose syntax he can describe only in the vaguest terms. Imagine that the study of French grammar in French schools was completely unsystematic, and then imagine asking a Frenchman about the word 'en'. But given an extensive French concordance, with the

possibility of translating everything in it, the uses of 'en' can be discerned and classified fairly accurately.

However, making a concordance of a Chinese text presents an awkward problem. Once again, the problem is how to represent the Chinese characters. A concordance in terms of the Chinese characters themselves is mechanically impossible. A concordance in terms of telegraphic code numbers is not very useful (unless one happens to be a trained Chinese telegraphist) because the whole point of a concordance is to present information so that it can be rapidly scanned. It is easy enough to add phonetic transcriptions to the code numbers, but these are so ambiguous that even a Chinese is usually unable to read a line of a concordance using phonetic transcriptions without the traditional characters.

It is hoped, however, that a useful concordance could be made out of a Chinese text in which every character is represented by its code number, and characters and phrases are additionally represented by equivalents given in a dictionary. In short, once the dictionary contains all the words and phrases on the lowest level of difficulty, a machine output like that reproduced on an earlier page could be the raw material for a useful concordance. The arrangement on the page would be more difficult than for an alphabetic language, but this could be solved at the cost of having either fewer lines to a page, or fewer characters to a line than the maximum of 24,

Conclusion

If this work were pursued for another year, it seems quite possible that some sort of machine translation of Chinese mathematical texts could be made. It might be necessary for the would-be reader to spend a few hours learning about the features of Chinese that lie behind the more glaring peculiarities of such translations. In particular, the arrangement of modifying phrases and clauses in Chinese leaves far more to the common sense of the reader than does the syntax of the various European languages. But common sense is what the reader of a machine translation must be relied on to possess in any case, while the greatest problem for any machine translation system, in the long run, is how to endow the program with a substitute for common sense. Such translations would certainly answer the needs of a mathematically-minded reader who had no easy way of getting Chinese mathematical texts translated properly.

One cannot leave this claim without noting two facts:

1. Mathematical texts are undoubtedly the easiest for machine translation to begin with, and other Chinese texts, even in closely related scientific fields, would present far more difficulties. But one must begin somewhere.

2. Any machine processing of Chinese texts involves a first stage in which someone with the telegraphic code at his fingertips is indispensable. If reading machines for alphabetic languages are developed soon to the point of combining flexibility with very low cost per word, one must hope that further development will enable them to handle an input as complicated as Chinese characters.

THE FRANKFURT KEYPUNCH CENTER

The Georgetown University Machine Translation Research Keypunch Center in Frankfurt was established to keypunch all text materials needed for the Georgetown group. The process of organization began in the fall of 1960. Georgetown rented 400 square meters of commercial space near the center of Frankfurt. Eighteen keypunch machines and twelve verifying machines were imported from the United States and installed. Production began in February 1961.

The keypunch instructions originally used at Frankfurt were revised late in the spring of 1961 and again in the fall. While these revisions interrupted production, the latest system was well received by the staff and solved all but minor problems.

Experience proved that, at the end of 3 months of training and production, a keypunch operator could be expected to punch seven to eight hundred cards per day, each card containing between 6 and 7 words. All work was verified. A verifier operator was able to produce 1100 to 1200 cards per day at the end of the same period.

The production of a keypunch operator with six to nine months experience ranged between 800 and 1000 cards; experienced verifiers produced between 1800 and 2000 cards. There was no significant difference in production between operators who knew the Russian language and those who had merely been trained to recognize the Cyrillic alphabet.

Final Production Figures for 1962

(Average cards per day, per person)

Month	Week	Keypunched	Verified
August	6-10	1023	1646
August	13-17	1029	1781
August	20-24	1110	1771
August	27-31	1055	1831
September	3-7	1190	1963
September	10-14	1234	1992
September	17-21	1249	2058

Despite early difficulties in recruiting and despite production slowdowns incident to retraining staff, more than thirteen million words were keypunched and verified. In addition to the Russian texts prepared for the Georgetown Machine Translation Research Project, materials in Dutch, French, German, English and Italian were prepared for Euratom.

In the spring of 1961 two native Russian speakers were engaged to compare the original texts with the printouts produced from the punched cards. The number of errors found by this type of checking proved to be insignificant and the checking was abandoned. It is noteworthy, however, that more than fifty percent of the misspellings found in the keypunched material were accurate copies of typographical errors in the original text.

The punched and verified cards were converted to tape on a locally available 1401 computer; the tapes were wrapped in lead foil and air-shipped to the Washington center. If the tapes were found to be satisfactory, the cards from which they were made were destroyed after six months.

FINAL PRODUCTION FIGURES FOR
THE FRANKFURT KEYPUNCH CENTER

Classification of Material	Number of Words
English translations of Russian texts	917, 785
Euratom Test	373, 000
Euratom French	333,710
Euratom German	136,701
Euratom Dutch	133,540
Euratom Italian	133,260
Internal Medicine	1, 277, 960
Celestial Physics	571, 140
Electronics	615,950
High Energy Physics	318, 065
Micro-Biology -	536,215
Crystallography	251,975
Geophysics	1,061,240
Solid State (Surface) Physics	698, 398
Soviet Propaganda	807, 055
Inorganic Chemistry	638, 225
Communicable Diseases	658, 945
Cybernetics	44, 100
Nuclear Physics	735,815
Physical Chemistry	467,765
Physiology	544, 950
Economics	312, 000
Organic Chemistry	345, 350
National Physical Laboratory	87,500
Miscellaneous Washington Test Materials	1,024,050

	Number of Cards
Pronouncing Dictionary	6, 760
Institute of Scientific Information	5, 180

THE PROGRAMMING SYSTEM FOR THE GAT

Serious programming design for the GAT began in the spring of 1958. By the end of 1959 a program was in operation on the IBM 705 II computer for the translation of Russian chemical texts into English, and a series of test runs was conducted on the computers of the Department of Defense at the Pentagon. This program was known as the Serna System.

As it became evident that IBM intended to discontinue the manufacture and operation of the 705 series, and as the need for a larger memory and faster access became acute, a study was undertaken for the conversion of the 705 programs into programs for another computer.

On the basis of the accessibility of different computers and on the basis of other technical considerations, the IBM 709 computer system was selected. Later the programs were redesigned and written for the IBM 7090 as the 7090 became more available than the 709.

In the course of the conversion, the 705 programs were completely revised and new ones were designed. This resulted in a completely new system. This program is known as the GAT program and also as the Direct Conversion Program. It is discussed in detail in the recent publication:

> Georgetown Automatic Translation; General Information and Operation Manual; John A. Moyne, Georgetown University, Washington, D. C., 1962.

Initially, computer time for assembling and testing this system was purchased commercially, but later, extensive time was made available to the project free of charge by the U.S. Air Force at the Pentagon, by the European Atomic Energy Commission (Euratom) at Ispra in Italy, and by the US Atomic Energy Commission at Oak Ridge, Tenn. A series of test runs were conducted in Italy in April and May 1961 by translating texts in chemistry, nuclear physics, cybernetics and so on, from Russian into English.

At present the GAT system is completely operative and can translate texts in several scientific and technical disciplines from Russian into English as a routine matter. Future work on this system should include further study of the linguistic formulations with a view to carrying out necessary revisions, as well as a study of certain more efficient programming possibilities such as the creation of a common library for input and output, a system of more efficient movement in the sentence, and the development, for all the routines of the system, of subroutines which may prove more productive.

DISCURSIVE DESCRIPTION OF THE SLC

The SLC system of programming for machine translation is described, from the user's point of view, in Georgetown University Occasional Paper No. 26; The SLC Programming Language and System for Machine Translation. This supersedes Occasional Paper No. 1; Manual for a Simulated Linguistic Computer--A System for Direct Coding of Machine Translation. (See also the paper on French-to-English Machine Translation.) The name "simulated linguistic computer" has been dropped; while it was descriptive to computer programmers, it was meaningless to any non-programmer who is interested in machine translation. The initials SLC are preserved as the designation of the system, but are no longer considered to represent words.

It seems fairly obvious that the process of automatic translation can be divided into two phases. First, a sentence or a larger group of words is looked up in a dictionary, usually word by word; this phase by itself is enough to produce a literally word-for-word translation in which every word of the input language has an invariable equivalent in the output language; the equivalent may consist of one word, more than one word, or even no words. To achieve a better result than this word-for-word translation, a second phase is necessary, in which not only a translation, but also certain linguistic information, is found in the dictionary for each of a special group of words; this group is usually a sentence. This material is processed all at once, so that the mutual relations of the input words can be discovered and used. One might say that, in this phase, each sentence is processed by a translation algorithm. Such an algorithm consists of a large number of steps. Some of these steps do not involve a decision, while others involve deciding, on the basis of observation of the particular sentence, which step to carry out next.

One more definition, that of the term 'item', will make it possible to describe a priori what kinds of steps a translation algorithm will contain. An item is whatever is brought from the dictionary to correspond to a single word of the input text. An item includes the original word in the input language to which it corresponds, at least one equivalent in the output language, codes which describe the syntactic and semantic properties of the input word, and, sometimes, codes which are not descriptive but executive, and which initiate sections of the translation algorithm whenever the items containing them are found in a sentence. Let us call the descriptive codes diacritics and the executive codes instructions. (This is not quite the same nomenclature as that used in the exhaustive description of the SLC system.) Then, each sentence of the input text (assuming the translation is effected sentence by sentence), will first be replaced by a series of items, one for each word of the input. This series of items can also be called a sentence, with no real risk of confusion.

The translation algorithm will include such steps as these:

1. The testing of the input word contained in an item, and therewith a decision as to which step to carry out next.

2. The testing of the equivalent output word contained in an item, with a subsequent decision.

3. The testing of any of the diacritics in an item, and a decision.

4. The testing of any of the instructions in an item, and a decision.

5. The alteration in the equivalent output word or words in an item. (Note that there is never an alteration of the input word, since the identity of the original word in the input text is an immutable fact.)

6. The addition, deletion, or modification of a diacritic or of an instruction in an item.

7. The deletion of an output item from the sentence.

8. The insertion of a new item at some point in the sentence. The new item will not correspond to any word in the original input; but this will often be a more flexible and satisfactory way of copying information into the sentence than the addition of information to an existing; item.

9. The replacing of an item in the sentence by a new item. This can be regarded as a deletion followed by an insertion at the same point, but it is convenient in practice to have the replacement available as a single procedure.

10. The moving of a group of one or more items from its previous position in the sentence to a new position. Such a rearrangement must involve the designation of three points in the sentence: the left and right limits of the group to be moved, and the point at which it is to be reinserted in the sentence.

11. The flagging of an item, the removal of a flag from an item, and the testing of an item for a flag. (A flag is a kind of diacritic which is temporary and simple, and which can be made to sum up the answers to a whole series of questions.)

Each application of a test or alteration will normally affect only one item in the sentence, and it is necessary to provide the algorithm with a means of addressing individual items. Several methods can be employed, but only the method actually in use will be described here. Whenever a section of the translation algorithm is initiated by an instruction, the item in which the instruction was found is automatically designated both as the source item (i. e. , the source of the instruction) and as the current item. The designation of an item as current means that until further notice, all tests and alterations carried out will apply to that item. The designation of an item as a source item is less important; its only significance is in the fact that a very convenient method is provided in the SLC programming system for making the source item current

at any time. The source item, then, is a starting point to which it is always easy to return.

Besides redesignating the source item as current, there are several ways of moving the designation of 'current item' from one item to another, and thus of shifting the attention of the translation algorithm from one item to another:

1. The source item can be redesignated as the current item.

2. The current item designation may be shifted one item to the right or to the left.

3. The current item designation may be shifted directly to the zero item. This is an artificial item which is added automatically to every sentence, and is considered to stand next to the left of the first item in the sentence, and also next to the right of the last item. The first item can be made current by going to the zero item and then one item to the right, and the last item can be made current by going to the zero item and then one item to the left. The sentence is thus like a chain in which the first and last links have been brought close to each other and joined by a special extra link.

4. Certain kinds of flags are restricted so that only one item at a time can be marked by such a flag. Any item so flagged can immediately be made current at any time.

5. Testing can also move the current item designation from one item to another. Most of the testings of an item which were sketched above exist in three forms:

   a) a simple test of the current item, which will yield a 'yes' or 'no' answer to the test and which will leave the current item still current whichever answer is obtained.

   b) A test of all items to the right of the current item, not including the current item itself or the zero item, to determine whether any of them satisfies the test. If none of them satisfies it, a 'no' answer is obtained and the current item remains current. If one or more of them satisfies it, a 'yes' answer is obtained, and the nearest item to the right which satisfies the test becomes the current item.

   c) an analogous test of all items to the left of the current item.

The linguistic algorithm can be programmed for a computer in a number of ways. If the algorithm has been described in a series of flow charts, the most obvious way to program it is to program each box in each flow chart independently, but with due regard to the proper conditional and unconditional transfers of control among the groups of computer orders corresponding to the boxes.

But once it has been observed that the algorithm is largely made up of a few dozen types of tests of items and alterations of items, a way of greatly reducing the amount of computer programming suggests itself. This is to program as many as possible of these tests and alterations in the form of closed sub-routines

Such a closed sub-routine is not directly built into the programming. If it were, there would be no alternative to repeating it at all the various points at which it is needed. A computer program, however, can be arranged so that any closed sub-routine can be stored once at a particular machine address, and a reference to this address will allow of the sub-routine's being used at every point where it is needed in the machine program.

A tremendous saving in bulk, and in programming time, can be effected if the maximum possible use of sub-routines is made in the linguistic algorithm. However, in ordinary programming, it takes one computer order to frame a simple instruction. (In the IBM 7090, the principle memory contains 32, 768 words; a word contains 36 bits; and a computer order occupies one word.) Suppose there are 512 sub-routines involved in the linguistic algorithm. Since $512 = 2^9$, it would be possible to specify a sub-routine by using only nine bits. Then a single word of computer storage, 36 bits, could express not only one instruction, but also a good deal of additional information in the remaining 27 bits of the word. A word organized in this way is here called a command.

In order to have a command of this kind properly carried out, an interpretive program is necessary. This program first locates, in the machine memory, the next command to be carried out. The command, a 36-bit word, is then broken into its components; one of these components is the 9-bit identification of the sub-routine to be used. There is also available a table of addresses, from which the interpretive program gets the address of the sub-routine in exchange for the 9-bit identification number. The components into which the remaining 27 bits of the command have been broken up are stored at special points in the memory, where they can be readily found and used by the sub-routines. For instance, one of the sub-routines may have the function of adding an instruction to the current item in the sentence. Any command which directs this to be done must include not only the identification number of the sub-routine, but also an indication of where the instruction to be added can be found.

In the SLC programming system as it stands, a single command can include indications of (1) the function which is to be carried out on the current item, (2) the location of a constant to be used in carrying out the function (e. g. an instruction to be added to the item, or a word in the output language which is to become the equivalent output word in the item), (3) the length of the constant, (4) the location of the command to be executed next if the present one includes a test to which a 'no' answer is yielded, (5) the location of the item which *is* to be executed next otherwise, (6) the location of the item which is to be current when the function is carried out; this may be the item which is already current, or the item next left, or the item next right, or the source item, or the item next left or right of the source item.

Pages 1 and *2* of Occasional Paper No. 26, The SLC Programming Language and System for Machine Translation are a "Brief Description of the Translation System", and they may serve as a continuation of this introduction.

It may be interesting to note here the size of the various programs in the system. The interpretive program which locates SLC commands and executes them occupies about 2, 200 words of memory, including the sub-routines for all the various functions. This is combined with routines that arrange the output from the dictionary lookup into sentences for processing, write out the translation of each sentence after processing, and monitor the progress of the linguistic algorithms if required; these routines occupy another 1, 200 words of memory. The linguistic algorithm for the present Russian-to-English system, coded for the SLC, occupies between six and seven thousand words of memory. Thus a total of about ten thousand words is occupied by the translation algorithm and the programs that implement it. About twenty thousand words of the computer memory remain free, and are used to store the results of each execution of the dictionary lookup program.

The overall translation process is carried out in cycles. Each cycle translates about two thousand words in one to two minutes.

The first phase of each cycle is the dictionary lookup; the program for this occupies about 1, 500 computer words. This program reads itself in from the system tape, and then reads a group of about 2, 000 words from the test and sorts them into alphabetic order. The program then reads the dictionary, and draws from it the items corresponding to the words of the text. The items are stored in the above-mentioned area of about 20, 000 words of memory, so that sorting them back from the alphabetic order of the dictionary to the original order of the text is a trivial task, rather than an affair of multi-tape and multi-pass merging such as it would become if larger groups of words were handled at a time. The last action of the dictionary lookup program is to cause the next program on the system tape to be read into the memory.

The next program is the translation program. It reads the translation algorithm in from the system tape, and then implements it. The translation program works its way through the items that were stored in the memory by the lookup program, and then writes out a sentence-by-sentence translation. Finally, the program rewinds the system tape. Then the next cycle begins. , .

There are two other considerable programs in the SLC system; the symbolic assembly program, and the dictionary updating program. These are briefly described on page 2 of Occasional Paper No. 26.

SAMPLE TRANSLATIONS

Here are some samples of the translations achieved by the Georgetown Machine Translation Project. (The reproductions of the translations are given at the end of this paper.)

Translation 1. This is from a French-to-English translation of a hundred-thousand-word text in physics. This is an "examined" text, that is, the text was checked to ensure that all of the text-words were in the dictionary, but no other preparation for translation was made. The form is that of the computer printout; it has been arranged to show both the original text (in which the French accents are rendered as post-positive numbers) and the English translation.

Translation 2. This is from a Russian-to-English translation of a forty-five-thousand word text in cybernetics. This is an examined text. The form of the sample is exactly that generated by the computer through the medium of a flexo-writer. (Two separate samples are given.)

Translation 3. This is from a Russian-to-English translation of a ten-thousand-word text in cybernetics. This translation was made at the same time and under the same conditions as the sample in Translation 2, except that this is a "random" text, and no preparations of any sort were made before it was translated. It is for this reason that certain Russian words appear transcribed rather than translated; they were not to be found in the dictionary as it stood at that time. (Two separate samples are given.)

These results have been achieved after less than eight years of subsidized research and after an expenditure of less than $1, 750, 000. 00. This budgeting has included heavy expenses for the hiring of computers; the lack of ready access to computers has made the work more arduous than it might otherwise have been. This budgeting also covers other areas of research which were less intensive perhaps, but of much theoretical and practical interest; these include the Comparative Slavic, Chinese-to-English, Arabic-to-English, and English-to-Turkish research, as well as all other research detailed in this report.

Various scientists, both American and European, who have seen our translations have stated that, while the English is not felicitous and the process of reading is slow, especially before the reader becomes used to the idiosyncracies of the output, the information conveyed is accurate, and the translation as it stands can be useful in scientific investigations. These statements can be documented for those who are interested, but since they were not originally made as public statements, they are not given here. One such statement, however, is on public record and may be repeated in this Report. It was made in 1960, and so refers to an earlier stage in the development of the Georgetown Automatic Translation System. (It is to be found in: Research on Mechanical Translation; Report of the Committee on Science and Astronautics, U.S. House of Representatives, Eighty-sixth Congress, Second Session, Serial d; June 28, 1960, Government Printing Office, Union Calendar No 895, Home Report No 2021, Page 48.)

"In brief, the translation can be understood by a chemist thoroughly familiar with the subject under discussion. A competent chemist can, from a study of the chemical formulas, reconstruct many of the chemical terms which appear to be a cross between a translation and a transliteration. With this background in chemistry, it is possible to work ahead slowly and systematically in order to make up for the obvious linguistic deficiencies of the machine translation and to grasp the principal train of thought. The reviewer was able to ascertain readily the essence of the paper which is contained in the paragraph running from 00134 to 00228.

"The translation of that part of the paper dealing with the experiments, if one disregards the rather poor "translation" of the chemical terms, is superior to that dealing with the theory. A chemist could repeat the experiments described in the paper by following scrupulously the procedures which are quite specific and intelligible.

"In summary, the reviewer believes that any good chemist, familiar with the subject matter and with a few peculiarities of the Russian language ("s" is translated "with", and "by" is translated "of") would find the translation usable.

"In his formal report to me, Dr. Weiss did not comment specifically on the economics of the translation. He informed me, however, that he estimated that it took him four times as long to attain a complete understanding of the significance of the paper as he would normally expect to require to read and understand an average man-made translation. "

The present output, then, can be read and understood, but slowly, by a person competent in the field. With a small amount of revision (which the continued refinement of the system should eventually render largely unnecessary), the translation can be made much more quickly comprehensible. Translation 4 is another section of the text represented in Translation 2, but here the unaltered machine output (on the left) is paralleled by a minimally revised text (on the right) in which the changes made have been underlined so as to be more readily identifiable. Human translators who have had a little training and some practice can revise a machine output in this fashion fairly quickly. (This material is a section of a fuller illustration given in the article Machine Translation and Language Data Processing, by L. E. Dostert; this article was originally published in: Vistas in Information Handling; Vol I, the Augmentation of Man's Intellect by Machine; Information for Industry Incorporated; Washington, D. C., 1963.)

$PAR LA FIGURE $FIG 11 REPRE1SENTE LE SCHE1MA GE1NE1RAL DE L APPAREIL UTILISE1 .
$PAR TE3TE DU MICROSCOPE .
NOUS NOUS SOMMES SERVI DE DEUX TE3TES ,.. CELLE REPRE1SENTE1E SUR LA FIGURE $FIG 9 EST DOTE1E D UN DISPOSITIF A2 CRE1MAILLE2RE QUI PERMET UNE COURSE VERTICALE DE L OBJET DE $FIG 60 MM ENVIRON .
CE DE1PLACEMENT IMPORTANT EST INDISPENSABLE SI NOUS VOULONS AMENER SOUS L OBJET UN CANON A2 IONS OU UN TUBE DE PULVE1RISATION CATHODIQUE DESTINE1S A2 DE1CAPER , DANS L ENCEINTE ME3ME DU MICROSCOPE , L E1CHANTILLON AVANT OBSERVATION .
POUR TRAITER THERMIQUEMENT L E1CHANTILLON , NOUS AVONS E1TE1 AMENE1S A2 UTILISER PAR LA SUITE UNE TE3TE ANALOGUE A2 CELLE DU MICROSCOPE A2 E1MISSION THERMIQUE (FIG. $FIG 10)
$PAR LE CHAUFFAGE DE L E1CHANTILLON PEUT E3TRE FAIT DE DEUX MANIE2RES ,.. SOIT PAR BOMBARDEMENT E1LECTRONIQUE (FIG. $FIG 10) , SOIT PAR CONTACT DIRECT AVEC UN PETIT FOUR EN STE1ATITE (FIG. $FIG 12) .
$PAR OBJECTIF (FIG. $FIG 13) .
LA PARTIE ESSENTIELLE DU MICROSCOPE A2 E1MISSION EST L OBJECTIF $FN (37) DONT NOUS RAPPELONS BRIE2VEMENT LE PRINCIPE .
LES FONCTIONS FOCALISATION ET ACCE1LE1RATION DES E1LECTRONS SONT ASSURE1ES SIMULTANE1MENT .
LES E1LECTRONS QUITTENT LA CATHODE SOUS UN ANGLE COMPRIS ENTRE $FIG 0 ET $FIG 90DEG ET AVEC UNE VITESSE $EXP *PHI-SUB-0 DE $FIG 0,1 A2 QUELQUES E1LECTRONS - VOLTS (CAS DES PHOTO - E1LECTRONS) .
LE PLAN IMAGE EST DE1FINI PAR L INTERSECTION DE L AXE OPTIQUE ET DU RAYON ISSU DU CENTRE DE LA CATHODE TANGENTIELLEMENT , AVEC LA VITESSE $EXP *PHI-SUB-0 .
$PAR CALCUL DE L OBJECTIF .
LE CALCUL A E1TE1 FAIT D APRE2S LES TRAVAUX DE $CAP SEPTIER $FN (37) SUR L OBJECTIF A2 IMMERSION .
NOUS NOUS SOMMES IMPOSE1 COMME CONDITION ,.. AVOIR UNE DISTANCE OBJET - WEHNELT DE L ORDRE DU CENTIME2TRE AFIN DE FACILITER L E1CLAIREMENT DE L OBJET .
IL S ENSUIT QUE CE SERA UN OBJECTIF A2 GROS TROUS , DE FAIBLE GRANDISSEMENT (NE1CESSITE1 DE LE FAIRE SUIVRE PAR UN PROJECTEUR) ET DE FAIBLE RE1SOLUTION .
$PAR PRINCIPE DU CALCUL .
ON APPLIQUE LA ME1THODE DE ((SUPERPOSITION)) POUR LE CALCUL DE L OBJECTIF ,.. ON DE1TERMINE LA RE1PARTITION DU POTENTIEL SUR L AXE EN COORDONNE1ES CYLINDRIQUES $EXP *(Z,R) D APRE2S LA FORMULE .
$PAR ON DE1TERMINE ENSUITE LA TRAJECTOIRE $EXP *R-SUB-Z D UN E1LECTRON ISSU TANGENTIELLEMENT A2 LA CATHODE , AVEC UNE VITESSE $EXP *PHI-SUB-0 = $FIG 0,7 EV .
SI CETTE TRAJECTOIRE COUPE L AXE A2 LA DISTANCE $FIG 250 MM (DISTANCE OBJECTIF - PROJECTEUR) , LA GE1OME1TRIE EST ADOPTE1E .
$PAR RE1SULTATS .
APRE2S PLUSIEURS ESSAIS , NOUS NOUS SOMMES ARRE3TE1 AU RE1SULTAT SUIVANT .
LES RE1SULTATS SONT REPRE1SENTE1S SUR LA FIGURE $FIG 14 .
$PAR POUR UNE VALEUR DIFFE1RENTE DE $EXP *PHI-SUB-0 , IL SUFFIT DE FAIRE VARIER TRE2S LE1GE2REMENT $SYM *A POUR QUE LA TRAJECTOIRE E1LECTRONIQUE COUPE L AXE A2 NOUVEAU A2 $FIG 250 MM .
$PAR LE CHAMP $EXP *E-SUB-0 , A2 LA SURFACE DE LA CATHODE , DONNE1 PAR LSSION E1TANT TRE2S FAIBLE , IL Y AURA UNE SOUS - EXPOSITION DE L E1CRAN FLUORESCENT ET DU FILM ET IL SERA DIFFICILE D APPRE1CIER LE CONTRASTE (COMME PENDANT L E1MISSION THERMIQUE POUR LES FAIBLES CHAUFFAGES) .

THE FIGURE 11 REPRESENTS THE GENERAL DIAGRAM OF THE APPARATUS USED .

HEAD OF THE MICROSCOPE .

WE HAVE USED TWO HEADS /COLON/ THAT REPRESENTED ON THE FIGURE 9 IS EQUIPPED WITH AN APPARATUS WITH TOOTHED BAR WHICH ALLOWS A VERTICAL COURSE OF THE OBJECT OF 60 MM. AROUND .

THIS NOTEWORTHY DISPLACEMENT IS INDISPENSABLE IF WE WANT TO BRING UNDER THE OBJECT A GUN WITH IONS OR A TUBE OF CATHODIC PULVERIZATION INTENDED TO SCOUR , IN THE SAME ENCLOSURE ITSELF OF THE MICROSCOPE , THE SAMPLE BEFORE OBSERVATION .

TO DEAL WITH THERMICALLY THE SAMPLE , WE HAVE BEEN BROUGHT TO USE CONSEQUENTLY A HEAD ANALOGOUS TO THAT OF THE MICROSCOPE WITH THERMAL EMISSION (FIG. 10) .

THE HEATING OF THE SAMPLE CAN BE DONE IN TWO WAYS /COLON/ EITHER BY ELECTRONIC BOMBARDMENT (FIG. 10) , OR BY DIRECT CONTACT WITH A SMALL FURNACE IN STEATITE (FIG.

OBJECTIVE (FIG. 13) .

THE ESSENTIAL PART OF THE MICROSCOPE WITH EMISSION IS THE OBJECTIVE (37) OF WHICH WE RECALL BRIEFLY THE PRINCIPLE

THE FUNCTIONS FOCALISATION AND ACCELERATION OF THE ELECTRONS ARE SECURED SIMULTANEOUSLY .

THE ELECTRONS LEAVE THE CATHODE UNDER AN ANGLE INCLUDED BETWEEN 0 AND 90DEG AND WITH A VELOCITY PHI-SUB-0 OF 0,: AT SOME ELECTRONS - VOLTS (A CASE OF THE PHOTO - ELECTRONS) .

THE PLANE IMAGE IS DEFINED BY THE INTERSECTION OF THE OPTICAL AXIS AND OF THE RAY ISSUING FROM THE CENTER OF THE CATHODE TANGENTIALLY , WITH THE VELOCITY PHI-SUB-0 .

CALCULATION OF THE OBJECTIVE .

THE CALCULATION HAS BEEN DONE ACCORDING TO THE WORKS OF //SEPTIER (37) ON THE OBJECTIVE WITH IMMERSION .

WE HAVE BEEN REQUIRED AS CONDITION /COLON/ TO HAVE AN OBJECT - WEHNELT DISTANCE OF THE ORDER OF THE CENTIMETER IN ORDER TO FACILITATE THE LIGHTING OF THE OBJECT .

THE RESULT IS THAT IT WILL BE AN OBJECTIVE WITH LARGE HOLES , SMALL ENLARGEMENT (REQUIRED TO FOLLOW IT BY A PROJECTOR) AND SMALL RESOLUTION .

PRINCIPLE OF THE CALCULATION .

ONE APPLIES THE METHOD OF ((SUPERPOSITION)) FOR THE CALCULATION OF THE OBJECTIVE /COLON/ ONE DETERMINE THE DIVISION OF THE POTENTIAL ON THE AXIS AS CYLINDRIC COORDINATES (Z,R) ACCORDING TO THE FORMULA .

ONE DETERMINES THEN THE TRAJECTORY R-SUB-Z OF AN ELECTRON ISSUING TANGENTIALLY AT THE CATHODE , WITH A VELOCITY PHI-SUB-0 = 0,7 EV .

IF THIS TRAJECTORY CUTS THE AXIS AT THE DISTANCE 250 MM. AN OBJECTIVE - PROJECTOR DISTANCE) , THE GEIOMEITRY IS ADOPTED .

RESULTS .

AFTER SEVERAL TESTS , WE HAVE STOPPED AT THE FOLLOWING RESULT .

THE RESULTS ARE REPRESENTED ON THE FIGURE 14 .

FOR A DIFFERENT VALUE OF PHI-SUB-0 , IT SUFFICES TO VARY VERY SLIGHTLY A FOR THE ELECTRONIC TRAJECTORY TO CUT THE AXIS AGAIN AT 250 MM.

THE FIELD E-SUB-O , AT THE SURFACE OF THE CATHODE , GIVEN BY LSSION BEING VERY SMALL , THERE WILL BE AN UNDER-EXPOSITION OF THE FLUORESCENT SCREEN AND OF THE FILM AND IT WILL BE DIFFICULT TO APPRECIATE THE CONTRAST (AS DURING THE THERMAL EMISSION FOR THE SMALL HEATINGS

but at them there were no feelings, they had not human capita. However later on they acquired a capita and became to annihilate people.

Interest to such topics completely understandable. The creation of an artificial reason - an idea to a higher degree capturing. We live into such ages, when most vain fantasies go over with the pages of the romans on the pages of scientific journals and books. The people created the submerged ships, invented the radios and the television effected launching of the artificial sputniks of Ground and the Sun. The possibility of a creation of an artificial brain from the object of pure fantasy transformed into the object of the most serious scientific discussion. (page 100)

An electronic man

Now as previously we heard concerning projecting of a bridge or the hydroelectric - station, and speak concerning the projects of a brain. True, the latter are discussed meanwhile only in a theoretical plan in view of the technical difficulties of their realization. But the possibility of surmounting of these difficulties in future, real practicability of an idea of a creation of an electronic brain detailed is substantiated by the authors of the projects.

We turn to which came out in 1956 g. to the collection "Automatic devices", a series of articles which were connected to by the questions of projecting of a brain. We study the project of a brain, aforementioned in one from them. The essence of reasonings : of its author D. G. Kalbertsona (see footnote) is reduced to following. (((Cf. "Automatic devices". The suppl. of articles, M., 1956, an article D. G. Kalbertsona "is Certain uneconomical robots"))).

A cybernetic system consists of elements, called by neurons, independently from their nature. In this system have elements, receiving action (receptors), reacting on it (effectors) and, finally, tying then and other (central elements).

The behavior of the man is defined by this, as connected to the help of central neurons external actions - the stimuli - to by our responding reactions. For example, a child cries - this influences on receptors - the organs of a hearing of a mother; its mother feeds - this a responding reaction. A worker at a machine tool obtains the assignment - the stimulus and performs it - a reaction or does not perform - an other reaction.

Receiving and reacting cells - receptors and effectors - unequal, at different people. Some people see and hear better, other - worse; some can move by extremities, other cannot (for example, as a result of a paralysis). But the contrast of people friend from friend is defined by the main way not by this, not by differences in themselves receptors and effectors, and by this, as they were connected between, i. e. by differences in central neurons, which unite receptors and effectors. From them depends the difference in the behavior of the people. One hurries on a help perishing, other saves; one coarse, other polite; one modest, an other boaster. All this depends on different connections of the stimuli and reactions, receptors and effectors, (page 101)

In the cybernetic apparatus these connections are defined by the program. It indicates, which actions and in which succession must be carried out after these or other actions.

Thus, the problem of a creation of an artificial man is reduced to to this, to collect together a sufficient quantity of receptors, effectors and, main, uniting their central elements and then to work out sufficient the complex programs, providing for such relations between by the stimuli and by reactions, which we have at people. Already now are created the robots, capable to perform certain functions of a man. More complex

function of dying of a man.

We will not speak concerning this, that the extinction of a robot from the blast of dynamite very similar little on this gradual process of an extinction, which is the death of a man. We turn attention only on, that in the given case a robot not only obtains a capability "to die" in 70 years, but acquires certain other properties, essential differing it from a man. For example, a man can rather safe for slip and to fall. And if falls a robot, in each cell which laid dynamite, then occurs a detonation and it explodes. True, possible to conceive the special apparatus, preventing the blast, or to replace dynamite by an other, more suitable substance. (page 139)But then, evidently, appear other properties, which again - such comes to compensate with the help of the special devices, etc.

A brain, as already spoke, contains an enormous quantity of neurons. This is necessary for the completion of its complex work. Tending to reproduce in a machine all its functions, a man creates in it such a property, as enormous dimensions, which also differs it from a brain. The brain of a man is placed in a skull, and an electronic brain would be equal according to dimensions to the colossal skyscraper. This circumstance, of course, cannot not mix such a machine to function, as a man. It is doubtful whether whether it would can play into hide-and-seek, even if and knew this play. Therefore the authors of the projects include into the number of the obligatory demands to the elements of a future artificial brain the small dimension of cells. But this already demand by no means of not a formal order. It is connected not to the form of the compound of cells, but to their substratum, to the properties of a substance, from which they consist. In order to satisfy this demand, necessary to change material, from which are prepared the elements of a constructed machine. Besides small dimensions these elements must possess and other properties, inherent to the cells of the brain: by a flexibility, by a capability to regeneration (to reduction), to the exchange of substances etc.

Shorter speaking, an artificial brain functioned just as natural, necessary, its dements - the cell functioned just as the elements of a natural brain - neurons. And this possibly only in this case, if all of the physical, chemical and also other properties of artificial and also natural neurons equal. Consequently, to create the machine, which functions, as the brain, necessary to make it from a substance, which possesses by these properties, - not from the electronic lamps or semiconductor elements, but from highly organized protein compounds, which form a natural brain.

kalbertson and the other authors of the projects of a brain incorrectly understand the relation of a function and a substratum, forms and contents. They proceed from that functioning of any system completely is defined by the relations of its elements, which we able to be in two states, and by different combinations of actions, which excite the corresponding cells. (page 140)Between by this the functions of a system depend not only from the relations between by elements, but on the properties of a substance, from which they consist, i.e. from their substratum. Therefore each from these elements can be not in two, and in much more diverse states.

In result in all said we must make a conclusion, that an electronic machine in a principle cannot completely replace the brain of a man. There is no whether here contradictions s kritikoi into the address of skeptics, which limit the possibilities of a development of a cybernetics ?

Contradictions not here. We speak in the given case not concerning the principal boundaries of understanding and possibilities of a human reason, and only concerning the definition of ways,, according to to which necessary to follow. , that impossible to make by by one method, necessary to attemp to make other. Upon the creation of the copy of this or other object exist two possibilities. In an one object can be accurately either

translation of books, yet they were based on completely other principles, than machines - translator with a constant algorithm, created into the first years of a development of a cybernetics. Before, for example, the machine begins to play into chess, it necessary to teach to this, on the other hand it will act not by means of the error of many variants of the play forward, but with the calculation of an experiment of the previous plays. A machine possible to teach and to a series of the different plays.

During further development of the structure the statistically - probable systems of a cybernetics will be ever more homogeneous, "biological", recalling the structure of a brain of the living creatures. In them will not be either the special measuring elements, of or the computing apparatuses, of or blocks of a memory. Upon its "a birth" they, according to - visible, nothing will not know to do/make and also them necessary will train.

Will in a parallel manner develop and regulated, or determinirovannye, systems, created for a completely definite problem. Among the latter a maximum development obtain so called extreme systems, a maximum of any index of quality. Being exposed in a book a theory is connected to this type of cybernetic systems.

In a book showed the community of the classification of cybernetic systems s samoizmeneniem ustavki, the programs, parameters, nonlinear characteristics either the structures and the possibility of extension of this classification on cybernetic systems from samoizmeneniem the algorithms, probable characteristics or the ranges of action of the system.

Upon the classification of systems, the conducting of analogies and analyses showed, that the systems, which utilize the principle of combined control, are most by complete regulated systems, which possess by maximum accuracy and by a high-speed operation. Upon combined control in a system is used simultaneously as control according to the degenerations, and control according to a regulated magnitude.

In a book established, that main two principles of control, being used upon the construction of the usual systems of automatic regulation (/ the principle of control according to the degenerations and the principle of control according to a regulated magnitude /), remain by the main principles of control and for cybernetic systems, and the consequently main methods of increase of quality (/of accuracy, high-speed operations /) are general as for usual, and for cybernetic systems of control. (o)--(o)--(o)--page 8 (o)--(o)--(o)--(o).

The first extreme system of a disconnected type (/ i.e. utilizing the principle of control according to the degenerations /) - the regulator of an angle of turning of blades of the hydroturbines - was effected by V. A. Bogomolov and by V. L. Beninym in a Institute electricians a ACAD.SCI. a UKR.SSR. An extreme system with closed back connection first described in a book G. Shteina "Regulation and alignment in parosilovykh apparati" (/ 1929 g. /). By one from the first systems of this type is and a regulator, which was worked out Naslenom (/ 2 /). First significant investigations of extreme systems with back connection were accomplished V. V. Kazakevichem (/ 1948 g. /), and then Draperom, Whether and Lainingom (/ 1953 g. /). The first extreme system of a kmobinirovannogo type - the system of regulation of the boiler on a maximum of a relation to par//topliv - effected in a Institute electricians AYU a UKR.SSR (/ 1958 g. /) under the manual of the author of this book.

In a book showed, that mathematical apparatus, necessary for the investigation of combined cybernetic extreme systems, comparative simple, often is reduced to the investigation of linear equations and that the difficulties of an investigation almost do not exceed the difficulties of an investigation of the usual systems of a regulirivani4, not possessing by the property of the device to a change of conditions of work. Thus, the

nonlinear and have a series of other lacks. In particular, they resolve the problem of regulation only in the relation to the main degeneration, according to to which was effected kompaundiruyushchaya connection.

Most complete there are the combined systems of regulation (16), in which is effected simultaneously as the principle of control according to to the divergence of a regulated magnitude, and the principle of control according to to the degenerations. The automatic systems of of the latter years all more frequently build as a system combined.

The many authors (/ and also Yuorbert Viner /) connect to cybernetic all of the systems of automatic regulation on this basis, that in them there are used managing signals. To us appears more feasible to call cybernetic more the specific range of systems.

Now already rather possible clearly to distinguish the systems of automatic regulation of dokiberneticeski of the period and a system cybernetic, which appeared after 1943 g. Study these and also other shows, that in the sense of principles of control the appearance of a cybernetics did not introduced nothing new. As the old, and latest cybernetic systems use indicated above the two main principle of control. In cybernetic systems we also find back connections according to a regulated magnitude (/or according to other internal coordinates /) or disconnected chonnechtions according to the Bain degenerations. The idea of combined control in a technical cybernetics in the same manner is fruitful, as in the systems of regulation of dokiberneticeski of the period. This position is by one from main in given to work.

It is naturally to prescribe a question:

If the principles of construction of the usual and cybernetic systems of automatic regulation some and the same, then in than consists their difference ((question mark)) a difference concludes, by the main way, in the problems of regulation, being decided by these and other systems, and in the method of effecting of back connections.

By the cybernetic system of automatic regulation possible to call a system, which was assigned for the decision of the new, more complex problems of regulation, than the classical problems of stabilization, program and follow regulation, (o)--(o)—(o)--page 13 (o)--(o)--(o)—(o). Back connections in cybernetic systems wide use the elements of logical action and often work in the regime of "a continuous search (/of oscillations /). In cybernetic systems questions, which were connected with control, by a generation, by the transformation, by transmission and by the procedure of managing signals, occupy the main place. Other questions (/ for example, questions transmissions energies /) have a secondary value. Here is a why cybernetics call also a science concerning the general laws of control.

We study a question concerning the problems of regulation, more detailed.

AUTOMATIC regulation up to the appearance of a cybernetics

With the time of invention of the first regulators a theory and the technology of automatic regulation developed as in the direction of expansion of a range of an application of regulators, and in the direction of increase of accuracy of a decision of various problems of regulation.

We enumerate the main problems of regulation.

The mentioned above regulators of a Ramelli, Polzunov and Uatta are connected to the most important and widespread class of systems of regulation, deciding the problem of stabilization of regulated magnitude on constant value. Still in a past century were

If compare words do not coincide, as a result this operation receives which anyone a number, but not 0. In this case occurs switchin on the following word of a dictionary, and so up to these being time, meanwhile upon a subtraction does not receive 0. 0 signifies, that a machine finding in a dictionary a combination, equal with given. Now necessary to know, which corresponds to it in by friend a tongue. Side-by-side with each word of a converted tongue is indicated the number of the cell, containing the corresponding combination of this tongue, on which convert. A when subtraction gives as a result 0, switching occurs already not on the following word of a dictionary, but on this cell of the second tongue, a number which is side-by-side with by the given word. (page 90) A combination of states of the apparatus "of a memory", concluded in this cell, gives upon an exit definite the alternation of holes and gaps on a blade, which is converted then on the usual tongue of letters. The when memory apparatus contains not the wholly words, but their bases and the grammatical indexes, then the machine seeks at first in the dictionary of the bases the maximum combination, which agrees with the first part of the given words, but then in the dictionary of the suffixes and also completions finds other its part. We permit, on the punched tape of a machine, converting from an english tongue on russian, perforated a word "letterless". In the dictionary of the bases turn out to be words "summer, Lett", "letter". A machine stops only on the latter, as far as it coincides with by a maximum part of the given in word and gives its translation: a letter, a learning, literacy. Then searches for a value which left to a part of a word - - less, indicating a negation, as a result of which on an exit is obtained a russian word "an ignorance", "illiteracy".

. . . . .

If compared words do not coincide, as a result of this operation a number is obtained, but not 0. In this case occurs switching to the following word of dictionary, and so on until a subtraction does not receive 0. 0 signifies that a machine finds in dictionary a combination equal with given. Now it is necessary to know, what corresponds to it in another tongue. Side-by-side with each word of a converted tongue is indicated the number of the cell containing the corresponding combination of the tongue, into which to translate. When a subtraction gives as a result 0, switching occurs no longer to the following word of dictionary, but on this cell of the second tongue the number of which is side-by-side with the given word. (page 90) A combination of states of the apparatus of the "memory" included in this cell, gives as an exit a definite alternation of holes and gaps on a tape, which is converted then in the usual language of letters. When the memory apparatus contains not the complete words, but their bases and the grammatical indexes, then the machine seeks at first in the dictionary of the bases the maximum combination, which agrees with the first part of the given words, and then in the dictionary of the suffixes and endings it finds its other part. For instance, on the punched tape of a machine converting from English tongue to Russian is perforated the word "letterless." In the dictionary of the bases turn out to be words "let", "Lett" and "letter." A machine stops only on the latter, as far as it coincides with a maximum part of the given word and gives its translation: letter, learning, literacy. Then searches for a value of the part of word left - less, indicating a negation, as a result of which in the output is obtained Russian word "ignorance", "illiteracy".

. . . . .

FURTHER DEVELOPMENT

The further development of the Georgetown Machine Translation System, and especially of the GAT, will take certain paths.

The Dictionary

There is no immediate need to increase the size of the dictionary, which contains entries derived from current literature in some five scientific fields, as well as basic materials which require fuller coding in some seven other fields. This coding will be completed.

A most important tool in this line of work is the considerable number of computer-produced concordances which have only just begun to be exploited.

These concordances will be utilized in improving the general system. In the dictionary certain fields of discourse which have already been keypunched will be coded and entered in the dictionary.

Other improvements in the dictionary have also been planned. A program for reviewing the clerical work and for eliminating the human errors in the coding of both the dictionary and the linguistic programs will be instituted.

Certain modifications which are designed to make the dictionary more compact and to increase the overall efficiency of the routines will be effected. These include making the length of the dictionary entry variable, arranging codes in terms of the frequency of use, grouping the codes in terms of their nature so that the Russian analytical codes will all be grouped in one area and the English synthesis codes in another, locating mutually exclusive codes in the same coding position, and reducing the length of certain codes.

In addition, when the SLC programming is used, all codes will be expressed in bit configuration, and this will decrease the time required for any operation which is directly related to the dictionary design, such as data-movement and code-testing; access time to the bit codes will also be decreased. The codes which have been interrogated in one routine and which are likely to be re-interrogated in a later routine will be retained in a position where they will be most easily available.

The Linguistic Program

The linguistic program is operational and is integrated with reasonably effective computer programs for the 7090. As part of the standard improvement procedure, adjustments will be made in both the linguistic statements and the computer routines.

The concordances will again be used in improving the linguistic operations, and in enlarging them. Specific studies will ascertain the distribution of

certain words, forms of words, semantic sequences, pattern sequences and - so on, in the texts. In particular, the area of semantic analysis will be explored

A system of semantic analysis will be evolved. The discussions in the Research Seminars in Semology have produced a number of suggestions which will aid in the development of a semantic analysis, at least of certain areas of the linguistic structure.

In proportion as semantic analysis becomes available, further elaborations in the English synthesis will bring the English output closer to idiomatic English. On the morphological, syntagmatic and syntactic levels, the English synthesis can be based on the existing output. A sizable obstacle, however, is created by the fact that the vocabulary selection in the target language must be almost entirely a direct function of the semantic control of the source language, and the English synthesis is largely a function of the vocabulary selection. Thus, the development of a semantic analysis of Russian is an important step toward a self-organizing synthesis of English.

The generalization of the existing syntagmatic and syntactic rules (about 35, 000 commands) will permit of expressing them more compactly, so that a decrease in volume of up to forty percent may be possible. Thus the introduction of a semantic analysis need not materially increase the size of the system.

In the procedure employed so far in the GAT, there has been a difference in the way that the source and target languages were analyzed. The target has been analyzed only in terms of dissimilarities from the source. The degree of dissimilarity from target to source was used as the chief guide in all the synthesis programs. The next important development must be the independent analysis of the target language and the evolution of a system of self-organization which will make the target language more consistent with itself. This will in turn lead to the development of a core system for Russian-to-English translation.

The core system for Russian-to-English translation can later be employed in other translation systems involving Russian or English and expanded as necessary. Such studies as the work in Comparative Slavic which will lead to the elaboration of a theory of core systems will be continued.

When machine translation becomes a commonplace, there will be a central linguistic analysis or core system into which the source language will be transferred, and out of which the target language will be transferred. Clearly, such a core system will produce an economy in the quantity of analyses needed when more languages than two are involved, and when each is used both as source and as target. In addition, such a core system will become the repository of information from which a more general theory of translation can be developed.

BASIC PRINCIPLES

The Georgetown Machine Translation Research Project has aimed primarily at producing a practical machine translation by practical experimentation. In attempting this goal, the research has followed certain principles. At the risk of restating material which may appear in the preface of this Report, these principles are given here.

1. Machine translation is essentially a linguistic problem. The process of translation has a language at its beginning and a language at its end. The corpus of language which forms the beginning is immutable to the translator. The corpus of language which forms the end may be mutable to a degree, but is necessarily less mutable in proportion as the translator is more accurate. Thus the source and target languages are the constants of any translation.

   A language is, amongst other things, an arbitrary system of conventional symbols. As such, a language has a systematized, but not a logical structure. The machine-translation researcher's task is to investigate this linguistic structure at all of its levels, and to describe it in a programmable form.

   Mathematical logic points to certain convenient processes in the description of language for machine translation. And the computer, as the basic tool by which machine translation is effected, necessarily imposes its limitations on the form of the description. But both logic and machinery are secondary aids in achieving a solution of the essential problem, which is the analyzing of the linguistic structure of the source language and the synthesizing of the linguistic structure of the target language.

   The facts of the language situation can never cede ground to considerations of what is more logical or of what is easier to program without a lessening of the efficacy of the machine translation system.

2. The best way to evolve a viable theory of machine translation is to begin with the facts of a specific problem of language transfer, to create a machine practice which will manipulate the languages so as to produce that transfer, and to distil a theory of machine translation from the actual practice of machine transfer. To begin with a general theory of machine translation and then to work toward a method is to be tempted to roughhew the facts of the language until they fit the preconceptions engendered by the theory.

   However, no unbridgeable chasm need be riven between the theoretical and the experimental approach to machine translation. On the one hand, it is difficult to imagine that a machine translation experiment can be conducted without any theoretical implications at all. On the other hand, it is pointless to conduct theoretical research which is not to be tested eventually in some experimental manner. The results of both approaches can and

should complement each other. Now the experimentalists forge ahead of the theoreticians, and now the theoreticians (having studied the data from the experiments) produce more concise and powerful formulations which can be used to rework the experimental findings.

3. The best way to evolve a method for machine translation is to translate a text by machine. The achievements of the early experiments may be very limited. The airplane at Kitty Hawk flew neither high nor far by present standards, but it flew. The first attempts at machine translation need not attempt everything which can be expected by machine translation later, but they must, however haltingly, translate.

4. An actual text, written for human perusal, is the only reasonable basis for research. Translating artificial texts, or texts which have been abstracted from actuality in any way is learning to swim without going near the water.

5. The first texts to be analyzed for the basic research are to be available both in the source language (the original text), and in the target language (a man-made translation). Source and target are to be analyzed independently as far as possible. Then a transfer system is to be constructed which will relate one analysis to the other. It is possible, of course, to analyze the source and target in terms of each other with a view to achieving, the most direct and economical translation, but this is a method which can commend itself only to those who never expect to translate except from that one source to that one target.

6. The linguistic research is to be text-focused. Thus, an item of vocabulary is generally to be entered in the dictionary only when it has actually occurred in a text. No word is to be entered simply because it occurs in a standard dictionary. Similarly, an item of structure is generally to be described only when the description is rendered necessary by features of the text. No structure is to be described simply because it is described in a standard grammar. Most particularly, no problem is to be solved because it might possibly occur, but only because it has occurred.

7. While the research is to be text-focused, it is not to be text-bound. Each statement is to be developed on the basis of the data observable in a text, but the total system of a class of structures is to be described exhaustively if an exhaustive description is reasonable in proportion to the frequency of occurrence of that type of structure in that type of text.

8. The development of the machine translation system is to be evolutionary and cumulative. A first text is to be prepared intensively, a machine translation of it is to be produced, and then the necessary adjustments to improve the translation system are to be made. Successive texts are to be treated in the same way, using the most recent version of the translation system. The translation system is to be consolidated and simplified as this becomes possible. In this way, the translation system is tested many times on texts of various types, and, as the new material is

incorporated, grows steadily until the time when further improvement is rarely necessary. However, further improvement must always be possible, and the addition of new items and the elimination of obsolete items must create no special problems.

9. In the elaboration of the machine translation system, the more obvious and frequent problems are to be dealt with first. An isolated difficulty may be overcome by means of an ad hoc solution. As more and more examples of that type of difficulty are encountered in later texts, however, it is to be expected that the multiplicity of ad hoc solutions will point to an ad omnia principle, which can then be incorporated into the system.

10. A new text of considerable length is to be used for each succeeding translation. A new text increases the size of the dictionary, offers new structural problems for solution, and avoids the possibility that the system will be well adapted to only one particular text. A long text of many thousand words provides the broadest possible prospect of the capabilities of the system. (It is of course, useful and interesting to retranslate the same text in each succeeding translation cycle, since this shows immediately the degree of improvement that has been achieved. But such a comparative study must remain purely ancillary if there is to be any important improvement in the system.)

11. There is to be no pre-editing of the source text to adapt it to the limitations of the translation system. There is to be no systematic post-editing of the output text so as to remove the effect of limitations of the translation system. The entire machine translation process must eventually be carried on inside the computer assembly. The final aim must be to produce a translation which requires only that degree of editing which is necessary in a translation produced by a human translator.

Ten years of experience indicate that these are the soundest principles on which to base a practical machine translation system.

SOURCE MATERIAL

The papers on which this General Report is based are listed below. After each title is given the name of the author, or the names of the authors. Because so many of the papers are the result of the collaboration of many people, however, it seems advisable to list also (in parentheses) the names of those who helped significantly in the necessary research.

Occasional Papers on Machine Translation (numbered and published)

1. Manual for a Simulated Linguistic Computer: A. Brown.

2. Key Punch Instruction Manual: A. J. Salemme (A. Boldyreff, E. Kalikin, J. Pyne and P. Toma).

3. The GAT Machine Dictionary: A. Boldyreff (E. Kalikin, D. Korn, J. Moyne, M. Pacak, M. Poltoratzky, J. Pyne, P. Smith, and M. Zarechnak).

4. The GAT Matrix and Its Information Retrieval: M. Zarechnak.

5. Morphological Analysis: M. Pacak (A. Boldyreff, D. Belmore, E. Kalikin, and M. Zarechnak).

6. The Dictionary Lookup: P. Toma.

7. The Exclusion Operation: J. Moyne and M. Zarechnak (A. Boldyreff, E. Kalikin, D. Korn, M. Pacak and J. Pyne).

8. Idiomatic Structures in Machine Translation: J. Moyne and M. Zarechnak (E. Kalikin, M. Poltoratzky, and M. Sushko).

9. Interpolation Routines: J. Pyne (E. Kalikin, M. Pacak and M. Zarechnak).

10. The Syntagmatic Analysis: E. Kalikin

11. GAT Syntactic Analysis: M. Zarechnak (J. Pyne).

12. Sentence Separators: D. Korn (J. Pyne and M. Zarechnak).

13. &14. The Resolution of Ambiguity of Noun Number and Case and The Case of the Adjective: E. Kalikin.

15. Lexical Choice: M. Zarechnak (A. Boldyreff, E. Kalikin, D. Korn, M. Pacak, M. Poltoratzky, J. Pyne, and P. Smith).

16. Verb Transfer and Synthesis: J. Moyne

17. English Synthesis Codes: P. Smith (N. Fargo, M. Pacak E. Pantzer and J. Rubin).

18. (Unpublished)

19. Particle Analysis: A. Salemme.

20. English Article Insertion: P. Smith.

21. Rearrangement: J. Moyne (E. Kalikin, M. Pacak, and M. Zarechnak).

22. The Morphological Abstraction of Russian Verbs: M. Pacak (A. Boldyreff).

23. Coding and Transfer of Czech Prepositional Structures: B. Chaloupka.

24. Machine Analysis of Russian Lexical Items in Organic Chemistry: P. Smith.

25. Rules for the Translation of Serbo-Croatian Prepositions Into English: M. Mellen.

26. The SLC Programming Language and System for Machine Translation: A. Brown. (This paper supersedes Occasional Paper No. 1).

27. Morphological Abstraction of Adjectivals in Czech: M. Pacak and H. Ulatowska.

Occasional Papers on Machine Translation (unnumbered and published)

Nested Structures in the Russian Language: M. Zarechnak and M. Mellen (K. Dekonsky, M. Poljak and I. Thompson).

A Fourth Level of Linguistic Analysis: M. Zarechnak.

Loci of Agreement and Government Structures in Slavic: M. Pacak.

Some Operational Solutions for Multiple Meaning in Machine Translation: M. Pacak and M. Zarechnak.

The Resolution of Lexical Ambiguity: M. Richman and I. Thompson.

Morphological Analysis of Polish Nouns: B. Henisz-Retman.

Morphological Analysis of Polish Verbs in Terms of Machine Translation: A. Woyna.

Georgetown Automatic Translation, General Information and Operation Manual: J. Moyne.

Terminal Reports on Machine Translation (unnumbered and unpublished)

The General Pattern of Linguistic Research in Machine Translation: M. Zarechnak.

The Machine Dictionary: B. Chaloupka.

The Transfer of Russian Words of Foreign Origin: B. Henisz and A. Boldyreff.

Morphological Analysis: M. Pacak.

Research Seminars on Semology: G. Trager (consultant).

Comparative Machine Translation Analysis in Slavic: M. Pacak.

French-to-English Translation Research: A. Brown.

English-to-Turkish Translation Research: R. Macdonald.

Research in Chinese: A. Brown.

The Frankfurt Keypunch Center: R. Heller

The Programming System for the GAT: J. Moyne.

A Discursive Description of the SLC: A. Brown.